GENE EXPRESSION PROFILING OF THE *nip* MUTANT IN *Medicago truncatula*

Brandon Lee McKethan, B. Sc.

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

August 2007

APPROVED:

Rebecca Dickstein, Major Professor
Juan E. González, Commitee Member
Robert C. Benjamin, Commitee Member
Samuel F. Atkinson, Chair of the Department of Biological Sciences
Sandra L. Terrell, Dean of the Robert B. Toulouse School of Graduate Studies

McKethan, Brandon Lee. Gene Expression Profiling of the *nip* Mutant in *Medicago truncatula*. Master of Science (Biochemistry), August 2007, 70 pp., 8 tables, 22 illustrations, references, 46 titles.

The study of root nodule symbiosis between nitrogen-fixing bacteria and leguminous plant species is important because of the ability to supplement fixed nitrogen fertilizers and increase plant growth in poor soils. Our group has isolated a mutant called *nip* in the model legume *Medicago truncatula* that is defective in nodule symbiosis. The *nip* mutant (numerous infections with polyphenolics) becomes infected by *Sinorhizobium meliloti* but then accumulates polyphenolic defense compounds in the nodule and fails to progress to a stage where nitrogen fixation can occur. Analysis of the transcriptome of *nip* roots prior to inoculation with rhizobia was undertaken using Affymetric Medicago Genome Array microarrays. The total RNA of 5-day old uninoculated seedlings was analyzed in triplicate to screen for the *NIP* gene based on downregulated transcript levels in the mutant as compared to wild type. Further microarray data was generated from 10 days post inoculation (dpi) *nip* and wild type plants. Analysis of the most highly downregulated transcripts revealed that the *NIP* gene was not identifiable based on transcript level. Putative gene function was assigned to transcripts with altered expression patterns in order to characterize the *nip* mutation phenotypically as inferred from the transcriptome. Functional analysis revealed a large number of chaperone proteins were highly expressed in the *nip* mutant, indicating high stress in the mutant prior to infection by rhizobia. Additionally, a database containing the information regarding the *nip* expression profile at both 0 days post inoculation (dpi) and 10 dpi were created for screening of candidate genes as predicted from sequence in the genomic region containing *NIP*.

ACKNOWLEDGEMENTS

I would like to thank the following people, without whom this work would not be possible:

- Dr. Rebecca Dickstein for the wonderful opportunity to work in her lab and for her constant encouragement through times of doubt.
- Dr. Catalina Pislariu for all her years of knowledge and experience and for her scientific critique.
- Dr. Juan González for his immensely helpful guidance and use of his laboratory.
- Natalie Gurich, Jennifer Morris and the González lab for their assistance with much of the RNA sample development.
- Viktoriya Morris, Yi-Ching Lee, Matthew Meckfessel, Dr. Laurent Coque, Dr. Harita Veershlingam, Bill Casady, Erich Spoor and all other members of the Dickstein lab for their help in all realms of science from plant care to PCR reactions.
- My family for years of support and encouragement.
- My wife Wendy who pushes me to succeed and has been behind me every step of this project.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

A17	Wild-type ecotype of *Medicago truncatula*
ABA	Abscisic acid
BAC	Bacterial artificial chromosome
C_t	A value calculated based on fluorescence of a real-time PCR sample. This value is generated when the DNA concentration crosses a threshold of fluorescence.
C90	The original sample name for the *nip* mutant
cDNA	Complementary DNA
cRNA	Complementary RNA
°C	Degrees celsius (centigrade)
DNA	Deoxyribonucleic acid
dpi	Days post inoculation
EST	Expressed sequence tag
GCOS	GeneChip operating software
latd	Lateral root organ-defective
LG1	Linkage group 1
mRNA	Messenger RNA
nip	Numerous infections with polyphenolics
PCR	Polymerase chain reaction
RIN	RNA integrity number
RMA	Robust multi-array average
RNA	Ribonucleic acid
RT-PCR	Reverse transcription PCR
Taq	Taq DNA polymerase enzyme
TBE	Tris-borate-EDTA
TIGR	The Institute of Genomic Research, now known as the J. Craig Venter Institute (JCVI)

CHAPTER 1

INTRODUCTION

1.1. Nitrogen Fixation and Global Nitrogen Usage

All living cells require nitrogen in order to survive, because nitrogen is a fundamental part of all amino acids/proteins as well as nucleic acids. Nitrogen, in its molecular form, is the most abundant atom in the air, comprising over three-quarters of the Earth's atmosphere. However, for the majority of life on Earth this form of nitrogen cannot be utilized for biological processes. Plants are particularly vulnerable to nitrogen deficiencies because they cannot actively seek out additional nitrogen sources due to their sessile nature. Therefore, plants often require external application of fixed nitrogen in the form of fertilizers in order to maintain proper growth. Almost all of this fixed nitrogen is generated via the Haber-Bosch process, an energy-intensive reaction that consumes approximately 1% of the world's energy supply (Smith, 2002). The nitrogen resources provided by the Haber process allowed for the worldwide increase in crop plant growth during the Green Revolution of the mid 20th century, which has contributed to increases in population growth that continue to this day (Borlaug, 2000). However, issues of global energy expenditure as well as possible side effects from nitrogen runoff have created a need to utilize other sources of fixed nitrogen for crop plants in order to ensure sustainability of the global food supply (Carpenter et al., 1998).

1.2 Bacterial Nitrogen Fixation and Plant Symbiosis

All identified nitrogen fixing organisms are prokaryotic. However, some nitrogen fixing bacteria form symbioses, with higher organisms such as plants, in which the bacteria provide the host organism with fixed nitrogen. In plants, most legumes are capable of this symbiotic

relationship, in which bacteria such as *Sinorhizobium meliloti* fix nitrogen for the plant in exchange for carbon energy sources. This symbiosis warrants study because of the possibilities of making it more efficient or transferring this capability into cereal crop plants. In third-world countries where the availability of Haber process generated fertilizers is less prevalent and/or too costly, a genetically modified cereal plant with the ability to fix nitrogen would provide a vast increase in food production in locations with poor soil. Additionally, use of such strains in first-world countries could lead to a decrease in the amount of fertilizer used, which would help prevent environmental issues with nitrogen product runoff, which can cause algal blooms and other problems such as the "Dead Zone" in the Gulf of Mexico (Rabalais et al., 2002).

1.3 The Model Organism *Medicago truncatula*

The model plant species, *Arabidopsis thaliana*, is incapable of forming nitrogen fixing symbioses, and therefore is not adequate for the study of nodulation. Additionally, the major crop species that are capable of nodulation, such as alfalfa and soybean, are less amenable to genetic studies due to issues with genome size, tetraploidy and inefficient gene transformation (Cook et al., 1997). Two legume models have therefore emerged; *Lotus japonicus* and *M. truncatula. M. truncatula,* or Barrel Medic, is a close relative to alfalfa (*Medicago sativa*), but because of its smaller genome, ability to self-cross and diploid nature, is more effective for use in classical genetic studies. Furthermore, *M. truncatula* has a short generation time, and can be transformed relatively easily using *Agrobacterium tumefaciens* and *A. rhizogenes* (Cook et al., 1997; Boisson-Dernier et al., 2001). *Medicago* is the dominant legume model system used in North America and Europe, with major efforts underway to sequence the entire genome (May and Dixon, 2004).

M. truncatula is a relatively small, day-length neutral plant that grows well in temperatures of around 22° C. It grows to maturity in 4-6 weeks and reproduces for 2-4 months depending on growth conditions such as the soil composition, fertilization and size of the pot the plant is grown in: if the plant is properly maintained in adequate soil conditions, it is possible for it to produce seeds for 6 months. It produces spherical seed pods that are approximately 1 cm in diameter. The seed pods are covered in thorns and become physically rigid and resistant to manual stress as they mature and are dropped from the plants. These pods must be crushed or pried open in order to remove the seeds themselves, which number 10-20 seeds per healthy seed pod. The seeds are kidney shaped and light brown in color and are a few millimeters in size.

1.4 Overview of Nodulation

The interaction between *M. truncatula* and *S. meliloti*, called nodulation, occurs when the bacteria come in contact with plant root hairs and initiate the infection of the root hair (Brewin, 1991; Brewin , 2004). The bacteria create infection threads which are elongated inward towards the inner portion of the root, until the bacteria reach the inner cortex where bacteria are released into host cells and begin differentiating into bacteroids. The plant also changes morphologically, forming a bump on the surface of the root called the nodule. In normal legume-*Sinorhizobium* interactions, the nodule grows into a rounded shape extending out from the root and becomes filled with the differentiated bacterial cells which can fix nitrogen. The nodule itself provides an environment with a decreased level of oxygen, as molecular O_2 is highly reactive and binds to the molybdenum core of the nitrogenase enzyme, inactivating it (Eady and Postgate, 1974; Orme-Johnson, 1992). This prevents the bacteria from fixing nitrogen in standard atmospheric conditions for oxygen. This forms the crux of the plant-microbe symbiosis: the plant benefits

because the bacteria provide a source of fixed nitrogen, and the bacteria in turn are provided with photosynthate for energy and a low-oxygen structure in which to fix the nitrogen.

The overall reaction equation for nitrogen fixation is:

$N_2 + 8H^+ + 8e^- + 16\ ATP > 2NH_3 + H_2 + 16\ ADP + 16P_i$ (Kim and Rees, 1992). The concern with this reaction is the very large amount of ATP required to break the dinitrogen molecule's triple bond: 16 ATP molecules are needed for each N_2. This heavy energy requirement makes nodule symbiosis an effective tradeoff for the bacteria, while the plant benefits by way of increased fixed nitrogen.

1.4.1 Initiation of Nodulation

The initiation of the nodule symbiosis involves a complex interplay of various chemical signals that prepare both symbiotic partners for nodulation. The process begins when the plant releases flavonoid compounds into the surrounding soil, which are perceived by the soil-living bacteria (Oldroyd and Downie, 2004). The bacteria then begin to produce various chemicals, termed Nod factors, in response to flavonoid perception. Nod factors are lipochitooligosaccharide molecules with various side-chain modifications that have a role in observed legume/bacterial symbiotic specificity (Schultze and Kondorosi, 1996; Schultze and Kondorosi, 1998; Perret et al., 2000). Nod factors from compatible species of bacteria are hypothesized to bind to the receptor complex NFR1/5 in root hairs which initiates an influx of calcium as well as membrane depolarization and swelling of the root hair tip (Cullimore and Dénarié, 2003; Oldroyd and Downie, 2004). This ion influx signal is then hypothesized to be transduced by the DMI1/DMI2 gene products, which then initiate Ca^{2+} spiking that activates the calcium-calmodulin dependent protein kinase (CCaMK) DMI3 (Mitra et al., 2004; Oldroyd and

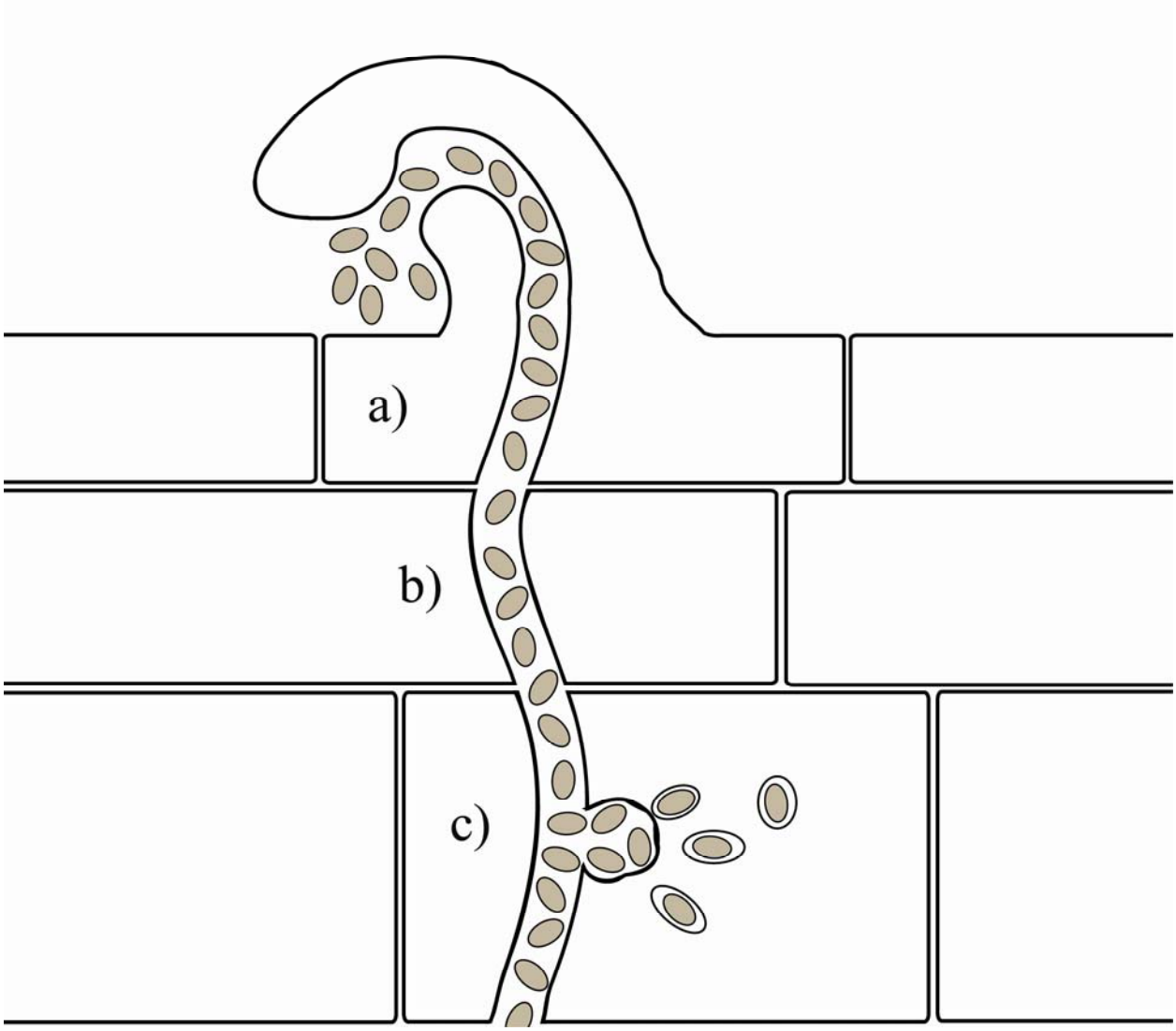
Downie, 2004). These genes products (DMI1/2/3) are also used in the signal transduction pathway for mycorrhizal perception, albeit with a different initiating surface receptor (Cullimore and Dénarié, 2003). In the nodulation pathway, DMI3 then directly or indirectly activates NSP2 and NSP1, a transcriptional regulator, which in turn initiates gene expression and increased cytokinin product in the root epidermal cells (Oldroyd and Long, 2003; Oldroyd, 2007). Cytokinin is transported from the epidermal cells to the root cortical cells by an unknown mechanism, where it activates the LHK1 receptor (Oldroyd, 2007). The LHK1 receptor is an activator of NSP2 that is both necessary and sufficient for nodule formation (Murray et al., 2007; Oldroyd, 2007). This NSP2 activation leads to the formation of the nodule primordium (Oldroyd, 2007).

1.4.2 Bacterial Infection

While the plant is tranducing the signals to begin cortical cell division and differentiation of the nodule structure, the bacteria begin the process of infection. The rhizobia initially attach themselves to the outer layers of the root hair, which initiates further root hair curling, leading to the formation of the distinct 'shepherd's crook' (Brewin, 1991; Brewin, 2004; Oldroyd and Downie, 2004). The bacteria that initiated the curling are then trapped in the shepherd's crook and initiate the infection thread. The bacteria divide in the infection thread and grow downward into the root hair and across the cortical cell layers. The thread itself is surrounded by plant cell walls does not interact with the plant intracellular space. When the thread has sufficiently elongated past the outer cortical cell layer and reached the newly divided cortical cells of the nodule primordium, the bacteria are then endocytosed into host cells forming what is known as the symbiosome. The symbiosome contains the bacteria in a plant membrane so that they do not

enter the intracellular space of the host cell. In the symbiosome, the rhizobia divide along with the symbiosome membrane and then differentiate into bacteroids, the nitrogen-fixing forms (Brewin, 1991, Brewin, 2004).

Figure 1 – Infection Thread Topology

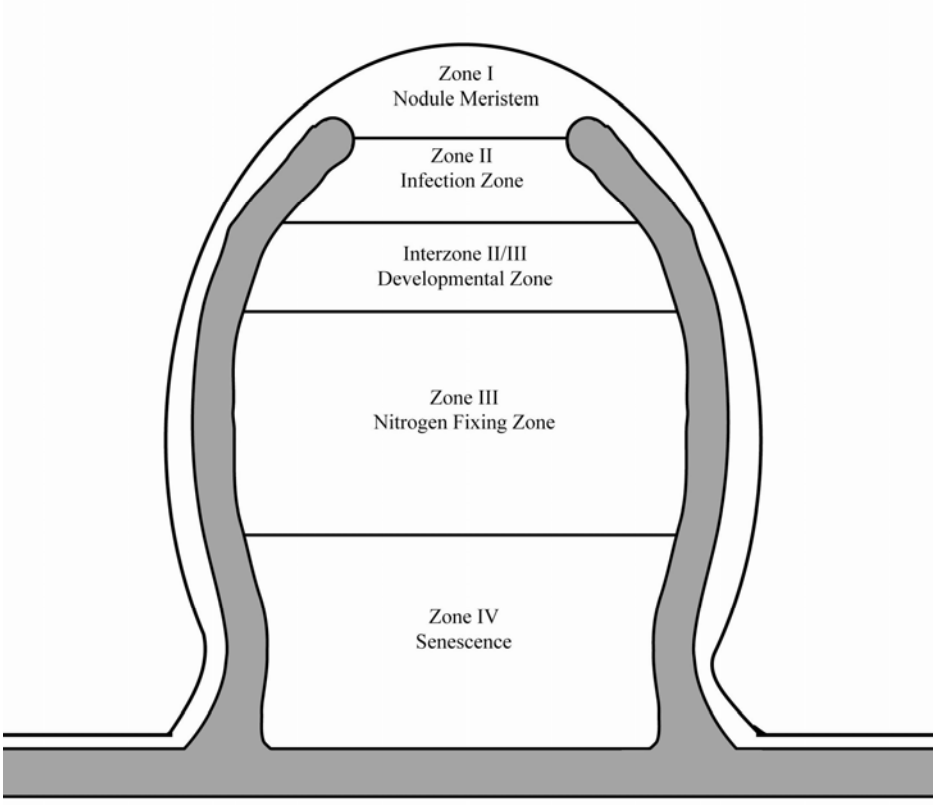


a) The bacteria initiate root hair curling, eventually forming the shepherd's crook and begin creating the b) transcellular infection thread matrix. Cells are c) endocytosed into the plasma membrane, which expand to form individual symbiosome membranes in the cortical cell layers. Figure after (Brewin, 2004).

1.4.3 Nodule Organogenesis

The differentiation of the nodule into various zones occurs as the organ develops. Figure 2 illustrates the different zones that occur in mature nodules. The top most portion of the nodule is the nodule meristem, which constitutes zone I and contains no bacteria (Vasse et al., 1990). The portion below is zone II, the infection zone, in which bacteria from the infection threads enter the nodule. In zone II, the symbiosomes and the host cells expand. Below zone II lies interzone II-III, in which the bacteroids complete their differentiate into nitrogen-fixing forms. The inner portion of the nodule below IZ II-III is zone III, the nitrogen fixing zone. A fourth zone exists in aging nodules, the senescent zone, in which aging bacteroids and host cells die approximately 5 weeks after inoculation (Vasse et al., 1990).

Figure 2 – Nodule Development Zones



The position of each zone of the nodule is shown: the grey portion around the other portion of the nodule represents the infection thread.

1.5 The *nip* Mutant

The screening of an ethyl-methyl sulfonate (EMS) generated mutant population of wild-type *Medicago* (A17 ecotype) identified the nodulation-defective mutant *nip*, or numerous infections with polyphenolics (Penmetsa and Cook, 2000; Veershlingam et al., 2004). The *nip* mutants were then backcrossed twice to minimize effects from secondary EMS mutations. Phenotypically, *nip* is characterized by relatively normal shoot development but markedly altered root growth, with shorter total root length and diminished length and number of lateral roots. Most notable about *nip* is its root nodule phenotype. When *nip* is grown in the presence of *S. meliloti* and in the absence of a fixed nitrogen source, the mutant becomes infected at a higher rate by rhizobia than wild type plants, but the nodules are small and rarely progress to a stage wherein the bacteria can begin nitrogen fixation. The infections typically halt before bacterial release into host plant cells (although there is a small amount of bacterial release), and nodules accumulate a large amount of brown polyphenolic pigments (Veershlingam et al., 2004). One hypothesis is that this apparent heightened defense response in *nip* and the related inefficiencies of infection thread development is caused by the mutant lacking the ability to differentiate the symbiotic bacteria from pathogenic strains. Thus, the *NIP* gene may play a role in friend vs. foe recognition in nodule formation (Veershlingam et al., 2004).

The *nip* plants' phenotype is also evident in its reproductive growth, as the seed pods produced by homozygous *nip* mutants are slightly smaller in size than those of wild type plants and typically contain fewer (5-10 total) seeds per pod than A17. Additionally, the seeds themselves are smaller in size and are less likely to germinate than are wild type seeds. This may be a function of lowered reduced nitrogen availability in the *nip* mutant.

1.6 The *latd* and *nip3* Mutants and ABA

Two mutants allelic to *nip*, called *latd* (lateral root organ-defective) and *nip3*, have also been identified by two separate research groups (Bright et al., 2005); Journet, personal communication). Complementation analysis has confirmed the allelic nature of the three mutants. The *latd* phenotype is more severe than *nip*; it has a very diminished root size as compared to wild type (Bright et al., 2005). The *latd* mutant forms nodules, some of which are infected, but the nodules are incapable of nitrogen fixation.

The hormone abscisic acid (ABA) is known to have some effects on lateral root development and the Harris lab demonstrated that the lateral root phenotype of *latd* can be partially rescued by ABA application (Liang et al., 2007). Interestingly, the nodulation phenotype of *latd* is not rescued by ABA. Similar results have been obtained with the *nip* mutant as well (C. Pislariu, personal communication).

1.7 Cloning by Expression Comparison

Analysis of gene expression can be used in conjunction with other methods, such as positional cloning, in the isolation of mutant genes. By comparing the expression of genes in the total RNA of both a wild type plant and a mutant, patterns of altered expression should arise, and if the gene in question is decreased in RNA abundance, it should be readily identifiable. In a sufficiently backcrossed mutant, secondary mutations should be minimized and background altered gene expressions should be decreased accordingly, which will assist in isolation of the single gene and corresponding pathways in question.

Previous work by Mitra et al. has shown that transcript based analysis of mutant *M. truncatula* plants can identify the single gene responsible for the particular mutation (Mitra et al.,

2004). Mitra et al. specifically showed that a single copy gene (*dmi3*) that had been deleted by fast-neutron mutagenesis could be identified by the lack of its transcript as compared to wild type using a DNA microarray. They furthermore showed that it could be used in a non-deletion mutant, identifying by transcript analysis the gene responsible for the *dmi2* mutation and an additional point mutation mutant in barley.

1.8 Microarrays and the Affymetrix GeneChip

Microarrays function in much the same manner as traditional Southern blots, in which DNA which has been fixed to a surface is treated with a labeled probe of interest, allowing sequences which are highly similar/identical to be hybridized to one another, marking the relative location on the fixed surface (such as a nitrocellulose sheet). The label used in most Southern blots is an incorporated radiolabel, but can also be a fluorescent label or other secondary label. The distinguishing feature of a microarray as compared to a Southern blot is that the array not only yields the location of the hybridization, but also has the additional function of interpreting the strength of the signal, allowing for comparisons of expression level based on the total amount of a given gene that hybridized to a certain sequence. Microarrays use a much smaller surface onto which a number of sequences or oligonucleotides (called probes) have been affixed, allowing a smaller sample volume to be washed over the entire surface, allowing many genes to hybridize simultaneously. For gene analysis, microarrays use the total RNA isolated from a particular sample and convert the mRNA by a reverse transcription reaction to cDNA. The more stable cDNA can then used to transcribe a biotin-labeled cRNA in the case of Affymetrix® chips (Affymetrix, 2004) (Affymetrix Corp., Santa Clara, CA, http://www.affymetrix.com). This labeled cDNA (comparison microarray) or cRNA

(Affymetrix GeneChip®) is then hybridized to the microarray, either simultaneously with a differently labeled cDNA product from a comparison sample, or as a single sample in the case of the Affymetrix microarray.

The Affymetrix GeneChip microarray allows for the comparison of gene expression with levels of absolute expression and typically larger probe sets than two-channel comparison arrays. Rather than hybridize two separately labeled comparison sets for relative expression, each RNA sample is hybridized individually to a separate GeneChip. Afterwards, comparative computational analysis is undertaken using normalized levels of expression, based on internal probes on each chip (Affymetrix, 2004). The Medicago Genome Array chip is an Affymetrix chip made for the analysis of *M. truncatula* gene expression. It also has all of the predicted genes of the corresponding nodulation bacteria, *S. meliloti*. A smaller number of alfalfa (*Medicago sativa*) genes are included as well. 50,900 total probes for *M. truncatula* exist on each chip, with 1,896 total for *M. sativa* and 8,305 for *S. meliloti* (Affymetrix, 2005). Of the probes for *M. truncatula*, 32,167 were generated from Expressed Sequence Tag (EST)/mRNA libraries and 18,733 were predicted from BAC sequence libraries from the International Medicago Genome Annotation Group (IMGAG). A total of 11 oligonucleotide pairs exists on each chip for each transcript (Affymetrix, 2005).

The other major factor distinguishing Affymetrix GeneChips from other microarray technologies is the use of multiple short (typically 25-mer) oligonucleotides to hybridize each transcript rather than longer stretches of cDNA (Affymetrix.com, 2007). This provides multiple benefits in data acquisition and replication. First is consistency between chips which is a function of the Affymetrix manufacturing methods. The chips are generated using a photolithographic technique where oligonucleotides are synthesized on-chip. As compared to

spotted cDNA arrays, this provides a high degree of reproducibility between different chips. This is central in how the Affymetrix system works: the inherent variability between spotted array cDNA microarrays makes them less suitable for chip to chip comparison, whereas Affymetrix arrays have higher consistency and can be used for running distinct, absolute value microarray experiments. Additionally, since the probes are synthesized on-chip via photolithographic techniques, Affymetrix microarrays can often fit more transcript probes per chip as compared to standard cDNA arrays (Affymetrix.com, 2007). Another benefit of oligonucleotide probes is that the software analysis of the GeneChip can interpret the individual binding and fluorescence of the multiple oligonucleotides (Affymetrix, 2001; Affymetrix , 2005). This provides a rubric for the specificity of the binding: if all the oligonucleotides of a certain probe have a signal, this is indicative of a specific binding event by the transcript, whereas if not all the signals are bound, this can be indicative of a non-specific binding event. Furthermore, each of these individual probe cells comprising a probe pair contains two separate oligonucleotides; one that perfectly matches the sequence and another with a single nucleotide mismatch (Affymetrix, 2001). By comparing the amount of signal present on both the perfect and the mismatched oligonucleotides, underlying noise and non-specific binding can be further removed from the signal.

1.8.1 GeneChip Operating Software

Affymetrix microarrays are conducted, recorded and analyzed with the freely provided software suite called GCOS, or GeneChip Operating Software. One of the major advantages of this software is that it creates standardized files from the initial experimental run of the chip, which can then be exported from the system and read by any other GCOS suite with the

appropriate library files. This allows users to conduct the experiment itself and the analysis at separate locations/times. Additionally, this allows the end-user of the data to generate any number of custom analyses and comparisons of their dataset using distinct, user-defined interpretation parameters. The program first analyzes individual GeneChip output files based on the levels of fluorescence detected on each of the multiple different oligonucleotides for each probe. By analyzing the individual fluorescence levels for different oligonucleotides, the software can determine whether the signal for that particular probe is the product of distinct binding of a specific transcript or simply non-specific binding. If not all of the oligonucleotides comprising a single probe have a signal, then the system will interpret that as indicating an "Absent" signal that is likely the result of background/non-specific binding. A corresponding p-value indicating the numerical basis for this Present/Absent call is also provided by the software. The software also generates the signal value from this analysis, which, based on the user's settings can alter the scaling parameters of the data. After the null-state (wild type) of the experiment has been analyzed individually, the experimental state (mutant) data is then analyzed similarly, but with the wild type as a baseline. This compares the internal Affymetrix control transcripts of each chip in order to determine the variance between the two chips, and then adjusts for those controls accordingly in assigning the m-value for each transcript. The m-value represents the log base-2 fold change between each transcript, indicating the quantitative level of expression change between the variable and control, in this case, mutant and the wild type. Along with the m-value, a separate p-value is also generated and a qualitative representation of that p-value is created, simply as "Change." This alphabetical modifier "Change" can either be NC, D, I, MD or MI, representing No Change, Decrease, Increase, Marginal Decrease and Marginal Increase, respectively.

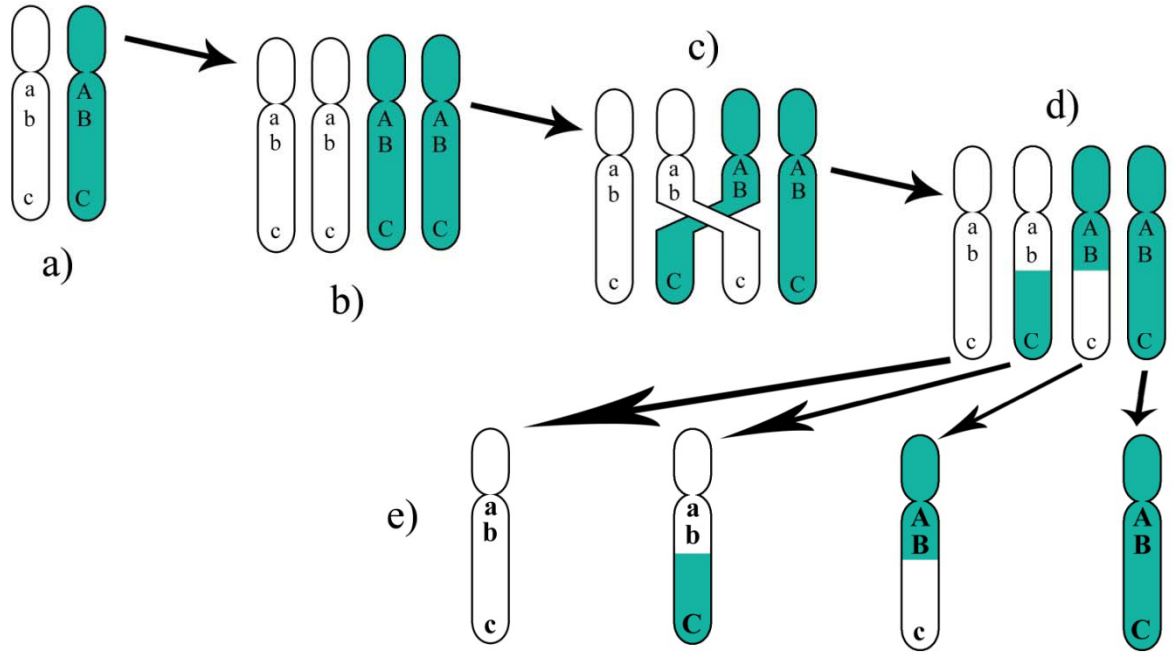
1.9 *Medicago truncatula* Genome Sequencing Project

The genome of *Medicago* is currently being sequenced as a collaborative effort between multiple groups at The Institute of Genomic Research (TIGR, now known as the J. Craig Venter Institue or JCVI), the University of Minnesota and the University of Oklahoma (Young et al., 2005; Medicago.org, 2007). The project is expected to be completed in 2008. Of particular importance to the cloning of the *nip* is the sequencing of chromosome 1, which is being conducted by Dr. Bruce Roe at the University of Oklahoma.

1.10 Positional Cloning of the *nip* Gene

The major effort towards identifying the location of the *NIP* gene in the *Medicago* genome is being conducted via a positional cloning experiment. The first portion of this positional cloning effort is being undertaken by genetic mapping of the *NIP* gene. This effort consists of taking the *nip* plant, originally called C90, and crossing the plant into a different wild type ecotype background in order to create a mapping population in which the mutant and wild type DNA come from polymorphic ecotypes. The C90 plant, having been generated from the A17 ecotype background, was crossed into the A20 ecotype. Plants from this cross were propagated to the F2 generation and screened for phenotype and recombinations between two markers flanking *NIP*. Figure 3 gives an overview of the process of recombination and how it allows us to distinguish genomic DNA of different ecotype backgrounds. The genetic area of interest is screened by using genetic markers for DNA length polymorphisms. Polymorphic regions in the genome, having changed over the genetic time separating the two ecotypes, allow us to distinguish between DNA of A17 (*nip*) and A20 (wild-type) in polymorphic loci using PCR or DNA sequencing. Multiple polymorphic regions, called markers, are then analyzed in

Figure 3 – Genetic Recombination

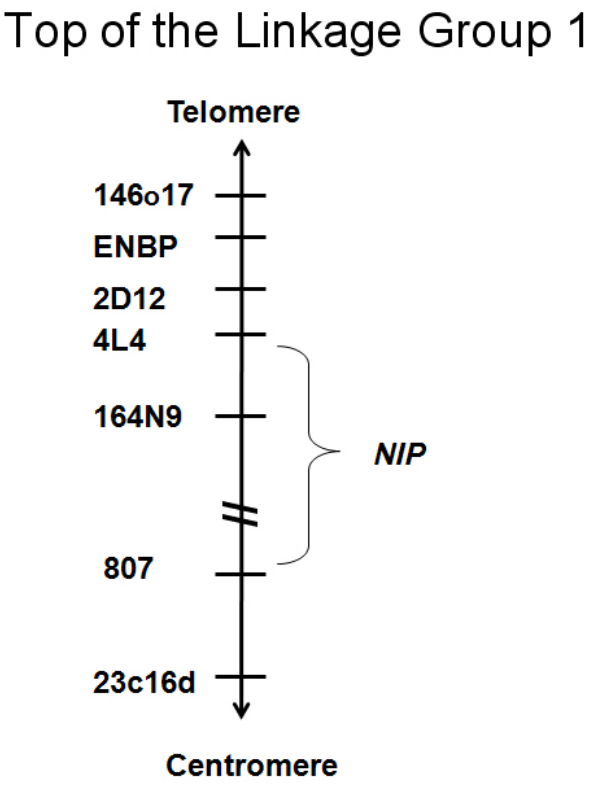


Crossing over of chromosomal DNA during recombination events. A heterozygous parent contains (a) one copy of the wild type copy of the chromosome (uppercase letters, teal color) and one of the mutant (lowercase letters, white color). During meiosis, the chromosomes are first duplicated (b), during which time a crossover (c) between chromosomes can occur. The resulting chromosomes (d) now comprise of part of the wild type and part of the mutant chromosomes. Four different gametes (e) containing different recombinations of the chromosomes are then generated. Genetic recombination allows us to generate plants with crossovers in the region near our gene of interest. By comparing results from genetic markers to the known phenotype of the plant, we can narrow down the region that contains the gene. In the case of *NIP/nip,* heterozygotes are indistinguishable from *NIP/NIP*.

locations spanning the putative location of *nip* and compared to the known phenotype of the plant, as ascertained when the plants were first grown in aeroponic chambers. Because *nip* is recessive, locations can be screened based on whether the DNA in that region of the chromosome matches the phenotype, such that, for example, regions containing heterozygous DNA in a *nip* phenotype plant can be reasonably excluded as not containing the *nip* gene. By analyzing markers at increasing closer intervals, the area in which *NIP* can possibly reside steadily decreases until the region is sufficiently small to begin analysis by a separate technique

such as standard DNA sequencing. Currently the *NIP* gene is localized to the upper portion of chromosome 1 (Linkage Group 1 or LG1) based on work conducted in the Dickstein lab by Heath Wessler, Viktoriya Morris and others.

Figure 4 – Genetic Map of *NIP* Gene

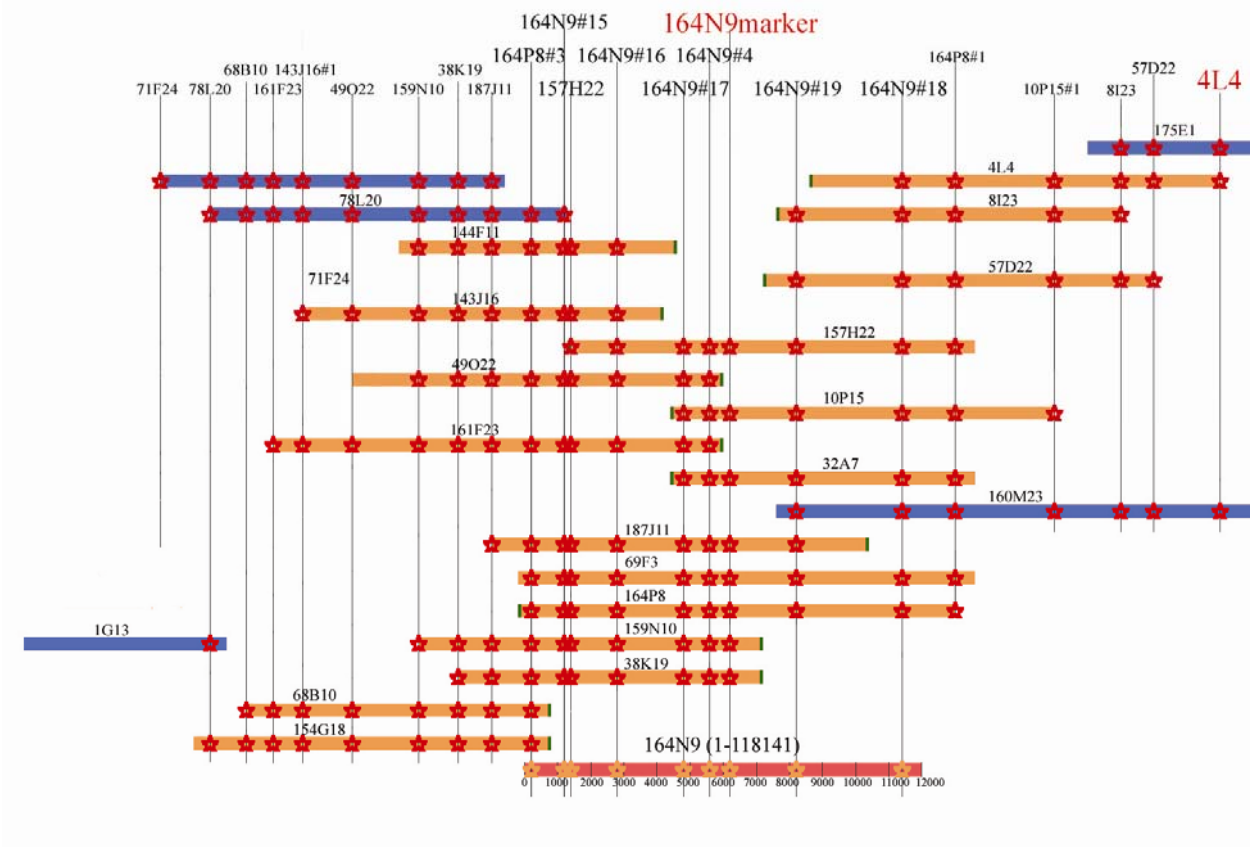


Genetic location of the *NIP* gene as determined by testing genetic markers in a recombinant population of *nip* and wild-type (A20). Each horizontal line represents the corresponding genetic marker tested at that location confirming the orientation of the gene. Image by Viktoriya Morris.

A second effort is being conducted in parallel via physical mapping to isolate the location of the *NIP* gene. This is being integrated with the genetic mapping. In comparison to genetic mapping, which relies on recombination frequencies (in the form of Morgans) to map the location of the gene, physical mapping attempts to piece together the bacterial artificial chromosomes (BACs) clones that comprise the area surrounding *NIP*. These BACs were

generated as part of the *Medicago* sequencing initiative in order to begin sequencing by the ordered clone approach. By piecing these fragments of the chromosome together, we can identify the physical area containing *NIP* and begin to search the sequence of this area as the results of the sequencing project become available over time. Additionally, this physical map construction assists in the generation of new markers for polymorphic loci to distinguish between the mutant and wild type DNA. This work is being conducted in the Dickstein lab by Yi-Ching Lee and others. Figure 5 shows the current status of the physical mapping of the *NIP* gene.

Figure 5 – Physical Map of Physical Area Surrounding *NIP*



Overlap between various BACs as arranged by PCR analysis: stars denote primer sets (markers) used to amplify regions for assembly of the map. The names of each primer set are noted by the corresponding line above the primers. Image courtesy of Yi-Ching Lee, modified to fit.

1.11 Goals for Expression Profiling Project

We have determined a number of goals for this project:

Goal 1: Create a list of up- and downregulated genes in *nip* according to biological function and biochemical context.

By analyzing the phenotype of *nip* based on expression changes in a biological context, we can establish a baseline for how the mutant is responding internally to the mutation. For instance, we hope to characterize the changes taking place in the mutant regarding increased basal levels of defense and stress related gene transcripts. This data should help us to understand the mutation from a biochemical standpoint and will assist in understanding the function of the protein encoded by the *nip* gene.

Goal 2: Obtain and analyze equivalent data from the third replicate of 0-Day *nip* vs. wild type as conducted by Dr. Catalina Pislariu at the Noble Foundation.

This third set of data should further remove false positives from the current pool of transcripts that were determined by the first two replicates' data. Triplicate microarray data will then be publishable in the context of a larger publication concerning the *nip* mutant. The most truncated list of data will also be helpful in removing putative transcripts that need to be tested against the *nip* genetic area.

Goal 3: Is *NIP* transcription downregulated in the *nip* mutant?

Our working goal for the data from *nip* expression is that the transcript responsible for the *nip* phenotype can be identified based on lowered levels of gene transcript as a function of increased transcript degradation as per Section 1.6. If the altered *nip* gene is selectively degraded in the mutant, then it stands to reason that the transcript levels will be markedly lower in the mutant. The difficulties that arise in this situation are the possibilities that a) the transcript

is expressed at very low levels in wild type plants as well as *nip* or that b) the transcript level is not sufficiently altered in the mutant. The former situation would be difficult to resolve as the transcript would not be distinguishable from more highly altered transcripts and additionally because the software/Affymetrix chip might not be able to even differentiate the signal from background noise. The latter situation makes identification by gene expression unlikely, as the undegraded transcript will not have an altered expression pattern and will be summarily passed over in the analysis. Should the data from the 0-day microarray analysis be insufficient to identify the *nip* transcript, then the data will still be useful in constructing an understanding of the *nip* phenotype from a biochemical and gene expression standpoint.

Goal 4: Establish a time-course analysis of transcription in the *nip* mutant.

The Affymetrix experiments conducted by Dr. Catalina Pislariu at the Noble Foundation included a microarray comparison between 10 DPI A17 and C90 plants. At this point in the time course of nodulation, the wild type plant has well established nodules and phenotypically *nip* plant can be differentiated from wild type based on the accumulation of brown polyphenolic compounds in small stunted nodules. This gives us further temporal information regarding gene expression as concerning gene transcripts that are present/absent in the nodule. By analyzing transcripts based on both their A17/C90 comparisons as well as data concerning the transcript abundance at two distinct nodulation time points, we hope to gain a better understanding of the mutation as well as possibly identify the *nip* gene itself, should it not be expressed before nodulation. Additional experiments using RT-PCR across further time points of 0, 5, 8, 10 and 15 dpi will be conducted on transcripts of interest indentified in the 0-day/10 dpi Affymetrix screen.

CHAPTER 2

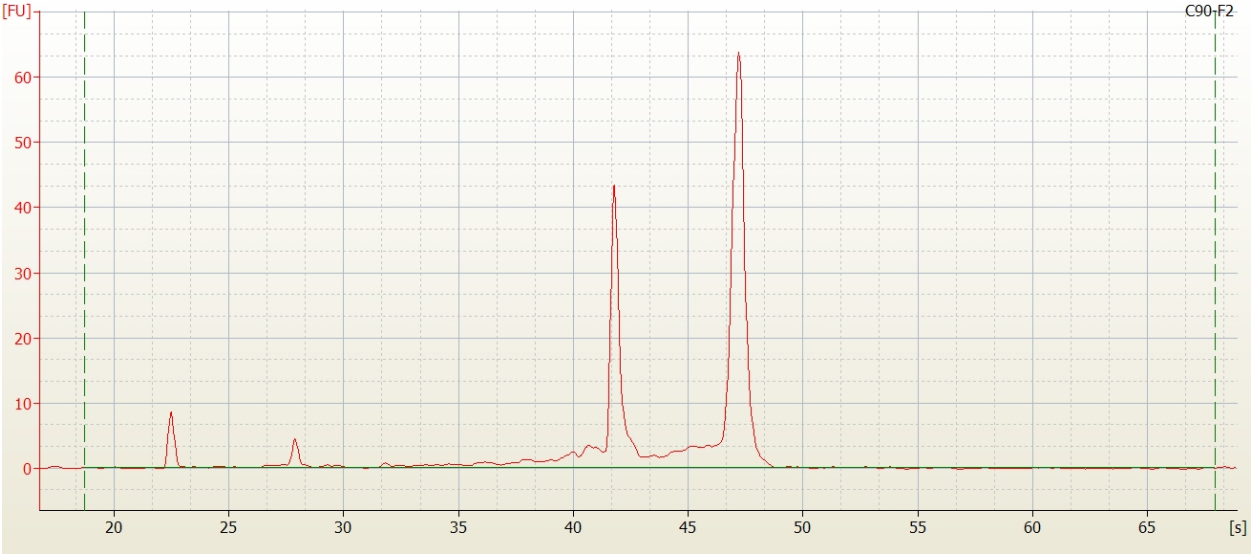
RESULTS

2.1 Isolation of RNA Samples

In order to assay the expression profile of the *nip* mutant, total RNA was extracted and purified from the roots of 5-day old uninoculated seedlings. The *nip* and wild-type control A17 seedlings were grown in aeroponic chambers on nitrogren-free media, thus starving them for reduced nitrogen. If inoculated with compatible rhizobia at this point, nodule development commences. This time point is termed 0-day. Two distinct biological replicates were conducted for the initial expression screen. Samples from these two different aeroponic chamber experiments were referred to as F and G, representing the first and second replicates, respectively. A total of 16 samples were extracted and tested for RNA integrity: 5 of A17-F, 4 of C90-F, 4 of A17-G and 3 of C90-G. This number of samples was generated to ensure that samples used on the microarrays had the highest RNA integrity. Of the samples generated, four were selected for further analysis using microarrays based on their respective sample readings for RNA Integrity Numbers (RINs), Real-Time genomic DNA amplification and reverse transcription polymerase chain reaction (RT-PCR) amplification of the Msc27 control gene.

RINs are a numeric calculation of RNA integrity based on electropherograms generated on microfluidics chips by the Agilent Bioanalyzer 2100 system (Meuller et al., 2004). Based on the amount of degraded background RNA and the 28S/18S ribosomal RNA ratio, a number between 1 and 10 is assigned to the sample, with higher numbers indicating better RNA quality. Samples above 9.0 are preferred for microarray use; a sample electropherogram with a high RIN is provided in Figure 6 along with RINs of samples used for microarray experiments (Table 1). Real-time RT-PCR was used as a second control for RNA integrity by amplification of the

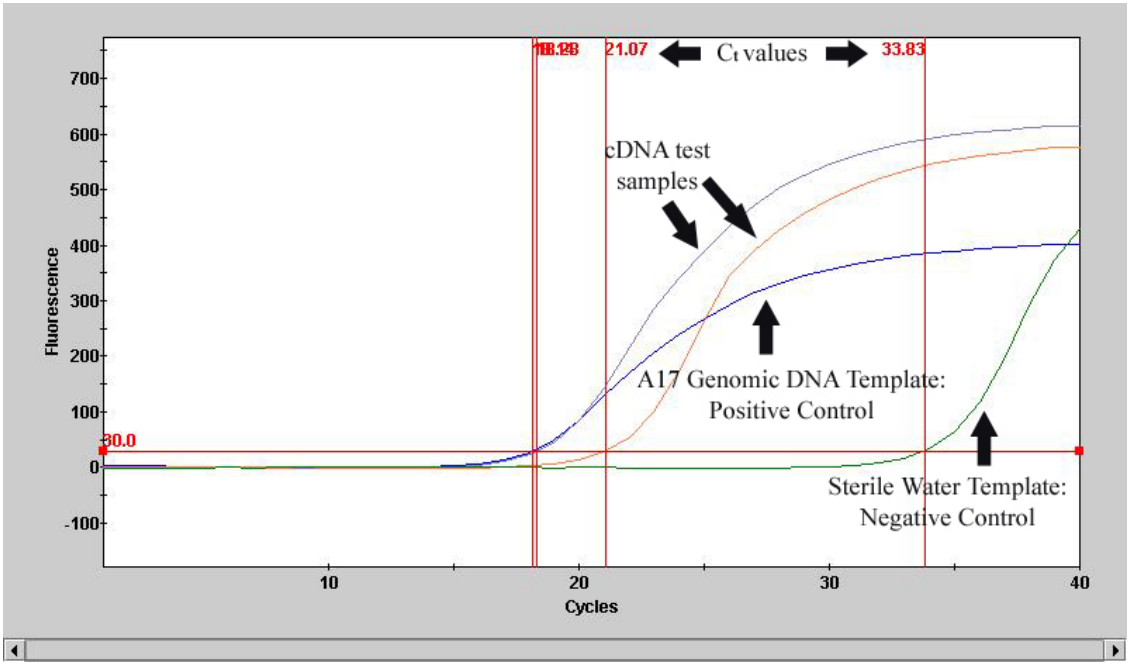
Figure 6 – Electropherogram of RNA Sample C90-F2



Electropherogram showing an RNA sample with a RIN of 9.4. The vertical axis is measured in [FU], or fluorescence units. This is proportional to the concentration of RNA passing through the microfluidics channel at that particular time in seconds, on the horizontal axis. The two major peaks at ~43 seconds and ~48 seconds are 18S and 28S ribosomal RNAs, respectively. The baseline amount of fluorescence, which is proportional to degraded RNA, is relatively low. Taken together, this indicates the sample is intact.

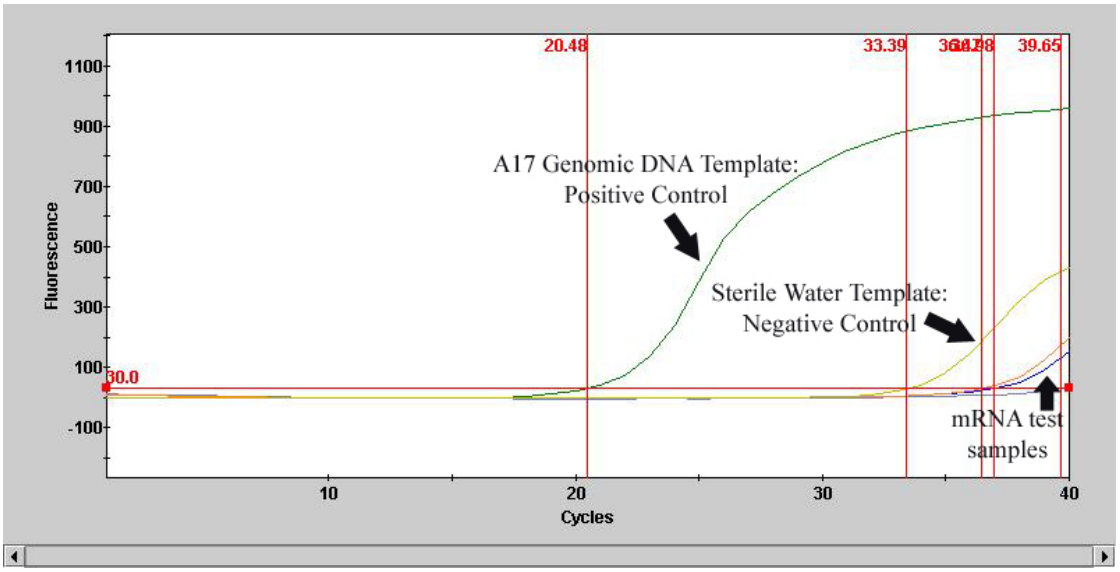
cDNA of a single control gene, Msc27 (Allison et al., 1993; Csanadi et al., 1994; Dickstein et al., 2002). This real-time RT-PCR provided a quantitative assessment of the amount of intact Msc27 cDNA present in each sample, which correlates with the level of RNA degradation in each sample. The value used for assessment was the C_t value, which, based the amount of initial fluorescence in the sample, provides a rubric for amplification from which the original amount of template in the sample can be estimated. The C_t value itself is a measure of the PCR cycle number when the amount of DNA that has been amplified crosses a given threshold. The lower the value, the higher the amount of template present in the sample, and correspondingly the more intact the RNA sample is assumed to be. A sample graph of real-time RT-PCR results is provided in Figure 7. Similarly, real-time amplification of the untreated RNA samples was conducted to confirm that all genomic DNA had been removed from the sample by DNase

Figure 7 – Fluorescence Readout of Real-Time RT-PCR Experiment



As the DNA in the sample is amplified, the amount of SYBR Green fluorescence (vertical axis) rises above a given threshold value (C_t). A lower threshold value indicates a larger amount of starting template in the sample. In this example, the two cDNA samples have similar C_t values and are below the positive control sample containing A17 genomic DNA as template. A sterile ddH_2O template is used to confirm the absence of any contamination in the amplification mixture.

Figure 8 – Fluorescence Readout of Genomic DNA Real-Time PCR Experiment



Real-time PCR amplification of samples with low levels genomic DNA, as noted by the higher C_t value of the samples as compared to the sterile water template negative control.

treatment. In this test it is preferable that the C_t value of the sample be close to or higher than that of the negative control, indicating that no significant amount of genomic DNA contamination is present. A sample graph of real-time PCR amplification of genomic DNA is noted in Figure 8. The information gleaned from these samples and their quality control data are presented in Table 1.

Microarray sample processing was done using the standard Affymetrix® protocol for eukaryotic RNA samples and run on the Medicago Genome Array chip (Affymetrix, 2004; Affymetrix, 2005) (Affymetrix Corp., Santa Clara, CA, http://www.affymetrix.com). This experiment was conducted by technicians at the University of Texas Southwestern Medical Center (UTSW) in conjunction with Dr. Juan E. González of the University of Texas at Dallas (UTD). All RNA quality control work was conducted at the González Lab at UTD.

Table 1 – Samples Used for Affymetrix Microarray Experiment

Sample ID	Original Sample Name	cDNA C_t value	RIN	Ratio of 28S/18S rRNA	Concentration (ng/ul)	Genomic DNA C_t value	Reference C_t value of ddH_2O
A17 #1	A17-F3	18.53	9.4	1.9	336.8	33.56	33.72
C90 #1	C90-F2	18.64	9.4	2.0	572.5	34.35	33.72
A17 #2	A17-G3	18.73	9.3	1.7	820.1	33.98	33.72
C90 #2	C90-G3	18.51	9.2	1.8	991.5	36.45	33.39

Sample ID – The ID number used for the identification of the data in the GCOS software. Original Sample Name – The original name used during the preparation of RNA samples cDNA C_t value – The C_t value of the real-time RT-PCR amplification of the Msc27 control gene from prepared cDNA; RIN – RNA integrity number; Ratio of 28S/18S rRNA – a comparison of size of the ribosomal RNA bands in the electropherogram of the RNA sample; Concentration – sample RNA concentration; Genomic DNA C_t value – the amplification value of the Msc27 gene (genomic copy) from total RN; Reference C_t value of ddH_2O – the respective amplification of a negative control containing only sterile water as a control.

Table 2 – Samples Used for Real-time PCR Confirmation of Microarray Data

Original Sample Name	cDNA C_t value	RIN	Ratio of 28S/18S rRNA
A17-F4	18.49	9.1	1.9
C90-F3	18.19	9.1	1.8
A17-G1	19.10	9.3	1.8
C90-G1	19.10	9.3	1.8

Sample ID – The ID number used for the identification of the data in the GCOS software. Original Sample Name – the original name used during the preparation of RNA samples; cDNA C_t value – the C_t value of the real-time RT-PCR amplification of the Msc27 control gene from prepared cDNA; RIN – RNA integrity number; Ratio of 28S/18S rRNA – a comparison of size of the ribosomal RNA bands in the electropherogram of the RNA sample.

2.2 A17 vs. C90 0-Day Data Screen – Initial Two Replicates

In order to determine which transcripts in the *nip* mutant were aberrantly expressed, data from the initial two microarray replicates were analyzed using Affymetrix's GeneChip® Operating Software (GCOS). Each replicate sample was compared internally using GCOS (A17 Replicate #1 compared to C90 Replicate #2) to account for internal controls on each chip that serve as baselines to normalize chip-to-chip variation. This corrects for chip fluctuations in signal to more accurately determine the change in expression between samples. Outputs from each comparison were combined using Microsoft Excel. The average $\log_2$ fold change (m-value) between replicates was used to sort the data as well as cull a shorter list of the most altered expression transcripts. The total number of transcripts identified with altered expression levels above a $\log_2$ fold change minimum is shown in Table 3.

Table 3 – Number of Altered Transcripts Identified by Replicate Data Screen

Minimum average m-value	Downregulated Transcripts	Upregulated Transcripts
2	129	95
1	559	342

Total number of altered *M. truncatula* transcripts with a minimum average $\log_2$ fold change (m-value) of *nip* compared to wild type in the first two replicates.

2.2.1 RMAExpress, Normalization and Chip Quality Assessment

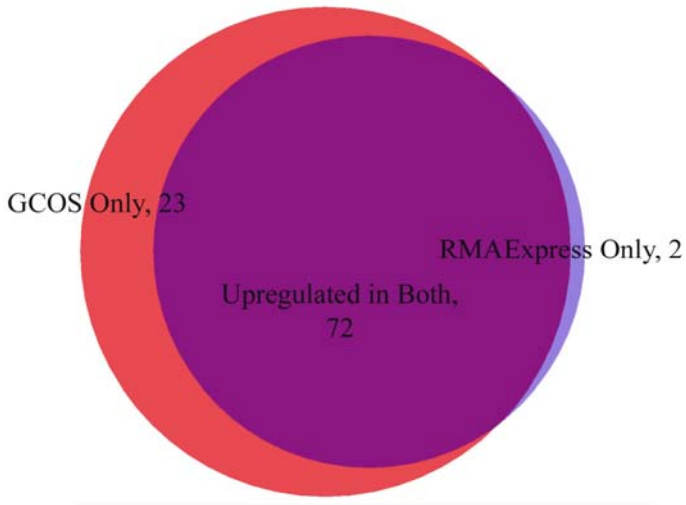
The program RMAExpress was used to calculate a separate set of probe signals based on the original data files of the different replicates. These .CEL files contain the raw output of the cell intensities of each probe cell on the GeneChip. The RMAExpress program calculates normalized data for each individual chip but does not provide direct chip-to-chip comparisons. The total data from the RMAExpress program was aggregated in Excel and compared by examining the ratios between wild-type A17 expression levels and *nip* (C90) levels. The samples were then culled based on the average ratio, with minimum average ratios of 2 and 4 corresponding to m-values of 1 and 2, respectively. These minimum values were chosen as cutoff points in order to focus on the most significant changes in expression. As the m-value falls below these cutoff points, the data becomes less precise and contains more false positives. These data were then compared to the previous output generated in GCOS (Section 2.2.1). The total number of transcripts with a minimum ratio of 4 or 0.25 is shown in Table 4. Common transcripts identified by both software analyses can be seen in Figures 9 and 10.

Table 4 – Number of Altered Transcripts Identified In A17 vs. C90 0 DPI Comparison by RMAExpress Software Analysis – Initial Two Replicates

Minimum Ratio of A17/C90	Downregulated Transcripts	Upregulated Transcripts
4 (Downregulated) 0.25 (Upregulated)	134	74

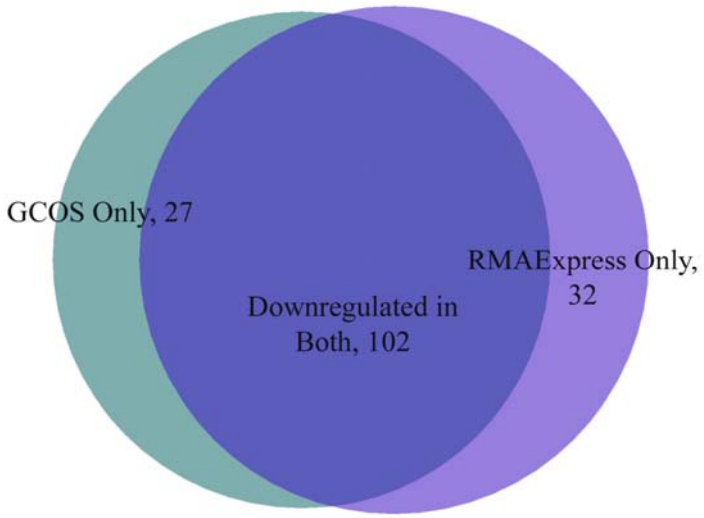
The total number of altered *M. truncatula* transcripts with a minimum average ratio equivalent to an m-value of 2 of *nip* (C90) compared to wild type.

Figure 9 – Common Upregulated Transcripts between Software Analyses – Initial Two Replicates



Total number of upregulated transcripts above an m-value of 2 (GCOS) or an equivalent ratio of A17 to *nip* of 0.25 (RMAExpress). Of the 74 samples in RMAExpress data set (blue) and the 95 found in the GCOS data set (red), 72 of those are the same (purple).

Figure 10 – Common Downregulated Transcripts between Software Analyses – Initial Two Replicates

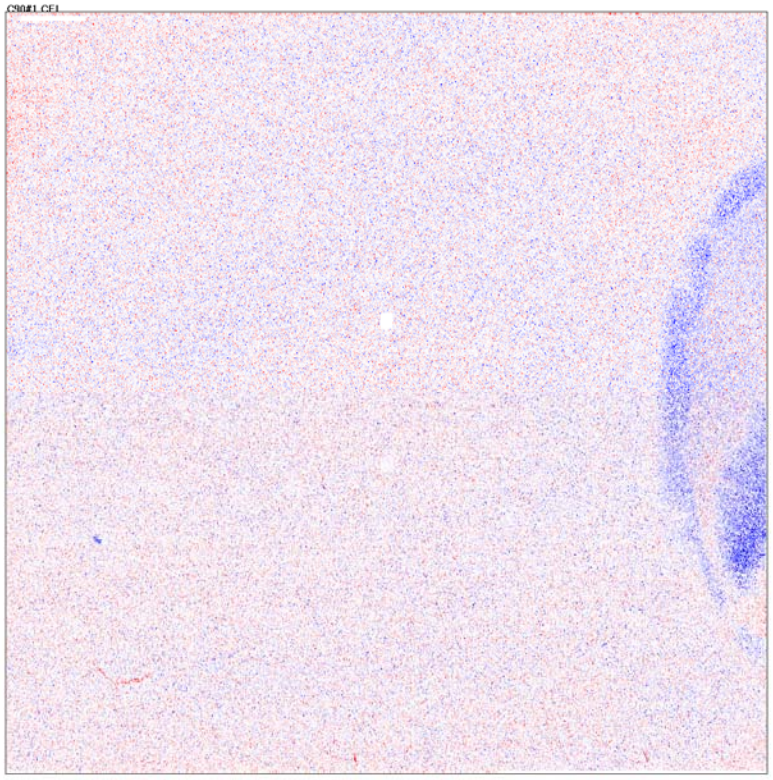


Total number of downregulated transcripts below an m-value of -2 (GCOS) or an equivalent ratio of A17 to *nip* of 4 (RMAExpress). Of the 134 samples in RMAExpress data set (purple) and the 129 found in the GCOS data set (aqua), 102 of those are the same (dark blue).

Additional quality assessment for each GeneChip experiment was performed by observing the patterns of altered expression on each chip using the program RMAExpress.

These residuals represent variations in expression at each probe cell on the chip as determined by the normalization and RMA calculations. Residual images provide a qualitative measure of the internal variations on each chip; any large scale patterns of expression levels represent possible inaccuracies in the data. Of the images generated in this manner, only the *nip* 0-dpi Replicate #1 (Figure 11) has any large scale variations, which are noted as negative values localized on the right side of the chip.

Figure 11: Residual Image of *nip* 0-dpi Replicate #1 from RMAExpress



Lower residual values on the chip are noted as blue dots, while higher residual values are colored red. A pattern of negative values (blue) are noticeable on the right side of the chip.

2.2.2 RT-PCR Confirmation of Microarray Data

To confirm the accuracy of the expression data obtained in the first two replicates, Real-Time RT-PCR of a number of transcripts was conducted. Six transcripts were selected; two that were upregulated, two that were downregulated and 2 having no change in expression in the

microarray analysis. Transcripts were chosen that had high expression levels in either A17 or C90 in order to be easily distinguishable by real-time amplification. Separately extracted RNA samples that originated from the same aeroponic chambers as the microarray test samples were used in real-time PCR analysis. These samples were also subjected to quality control experiments to control for RNA integrity by using real-time RT-PCR of the Msc27 transcript (Table 2). This allowed for a comparative baseline of expression between different samples. The data shown in Table 5 illustrates the comparison between the average fold change obtained by Affymetrix array to the average change in C_t value between A17 and C90 replicates by real-time RT-PCR. Both changes are expressed as $\log_2$ fold change in expression from A17 to C90. The results showed good correspondence between microarray and real-time RT PCR, thus validating the microarray data. The scale of both C_t values and m-values is measured as $\log_2$ fold change in expression from the wild type sample to *nip*.

Table 5: Comparison of Transcript Expression between Microarray Data and RT-PCR Amplification of Probes

Transcript (Probe Number)	**Log Fold Change Average (Affymetrix)**	**Log Fold Change Average (Real Time RT-PCR)**
Albumin 1 (Mtr.42382.1.S1_at)	-3.6	-4.875
MYB transcription factor homologue (Mtr.37886.1.S1_at)	-3.45	-2.51
Phosphofructokinase (Mtr.48868.1.S1_at)	NC	0.005
Thiolprotease (Mtr.25991.1.S1_at)	NC	-0.375
17.9 kDa heat shock protein homologue (Mtr.12214.1.S1_at)	4.55	7.375
17 kDa heat shock protein homologue (Mtr.39929.1.S1_at)	4.55	6.795

Validation of the GeneChip experiment by RT-PCR amplification of transcripts using SYBR Green fluorescence to signal for total cDNA amplification. RT-PCR data are expressed as C_t values while the Affymetrix data expressed as M-values; each of these indicates a $\log_2$ fold change in expression.

2.2.3 Bioinformatic Analysis of Downregulated Transcripts

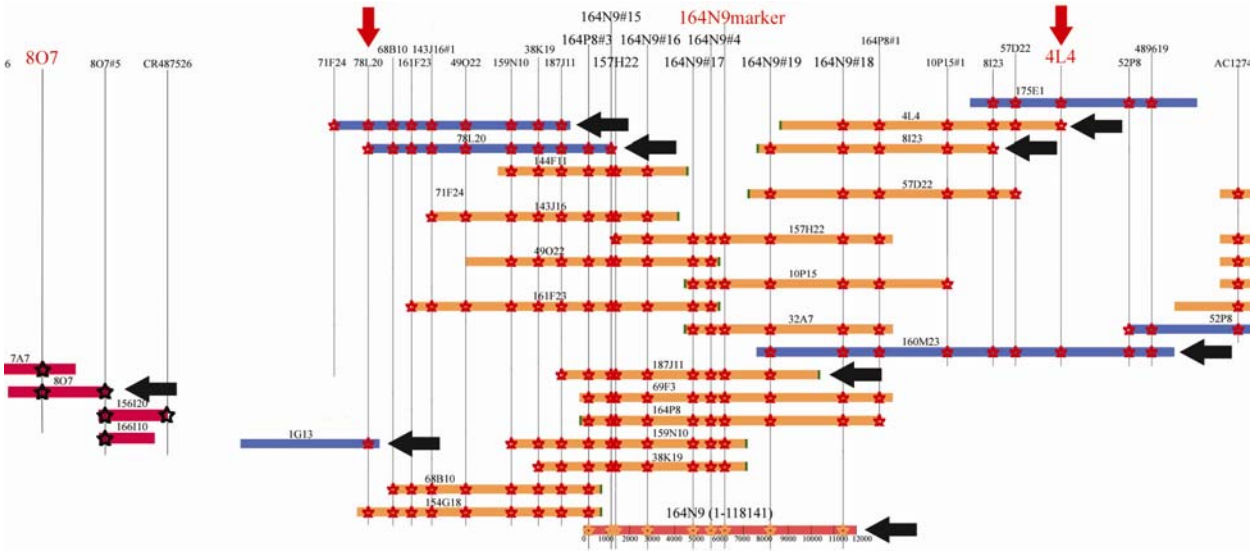
Based on previous work, it was hypothesized that the *nip* transcript might be identified based on the decreased expression of its mutant transcript (Mitra et al., 2004). Mitra et al. showed that point mutation genes in *Medicago* and barley that resulted in the preferential degradation of the mutant transcript could be identified by their lowered expression levels on microarrays. Therefore, the most highly downregulated transcripts determined by the microarray data were analyzed to determine if they could possibly be the *NIP* gene. The downregulated transcripts identified in Section 2.2.2 were screened first based on their location in the *M. truncatula* genome. The *NIP* gene is known to lie at the top of chromosome 1 (linkage group, LG1). Information available for each transcript such as the Bacterial Artificial Chromosome (BAC) clone containing the gene's location relative to genetic markers assisted in identifying which transcripts could not be *NIP*. Sequencing of the *Medicago* genome is being conducted using an ordered clone approach, and the BACs used in this approach are assigned to specific contigs and chromosomes, the data for which is available online (http://www.medicago.org/genome/). Approximately 35% of the transcript probes on the Medicago Genome Array chip were predicted from BAC sequences and thus are annotated with the appropriate BAC (Affymetrix, 2005). These can be searched against the medicago.org database to determine their genomic location. Alternatively, for probes that were predicted based on Expressed Sequence Tag (EST) data, genomic localization can be determined from sequence homology to BACs present in the National Center for Biotechnology Information (NCBI) nucleotide database. In these cases, the Basic Local Alignment and Search Tool (BLAST) function was used to compare transcript sequences against the NCBI database of *Medicago truncatula* genomic sequence. This allowed for localization of transcripts to certain

BACs and determination of their chromosomal location. By these methods and in conjunction with data from the third replicate microarray, all but 15 of the most highly downregulated transcripts were eliminated as candidates for *NIP*.

2.2.4 PCR Transcript Testing Against Physical Interval

Transcripts that could not be localized on the *M. truncatula* genome using bioinformatics were then tested against the known genomic area containing *nip* by PCR amplification. On-going work in the Dickstein Lab has identified BACs representing the minimum tiling path of the physical interval containing *NIP* (Figure 12, Figure 5). These BACs were used as templates in PCR amplification of each transcript. DNA from A17 and from another wild-type ecotype, A20, were used as positive controls in the PCR. This ensured that the transcript could be amplified from genomic DNA, because many of the transcript primers were based on mRNA sequence and therefore might span introns, interfering with the amplification. Electrophoresis gel results of PCR amplification of transcripts from the BACs are shown in Figure 13. Each sample contains amplification products from 9 different BACs covering the interval. Samples The results show that each PCR set contains 1 positive control (A17 genomic DNA) and a negative control (sterile water template) or as having two positive controls (A17/A20) and the negative control is not shown. The nine BACs used in each sample are (in alphabetical order): mth2-160M23, mth2-164P8, mth2-187J11, mth2-1G13, mth2-4L4, mth2-71F24, mth2-78L20, mth2-8O7 and mth2-8I23, unless otherwise noted. The test name and corresponding internal GeneChip name are included in each panel.

Figure 12 – BACs Used to Represent Genomic Interval of *NIP*



Overlapping BACs in the physical interval containing *NIP*. The *NIP* gene is thought to be in the DNA contained between the 78L20 and 4L4 genetic markers, which are marked by red arrows. The BACs used represent the minimum tiling path between 78L20 and 4L4. BACs used in the PCR amplification experiements are marked by black arrows. Original image provided by Yi-Ching Lee.

Figure 13 – PCR Amplification of BACs Comprising Minimal Tiling Path Containing *NIP*

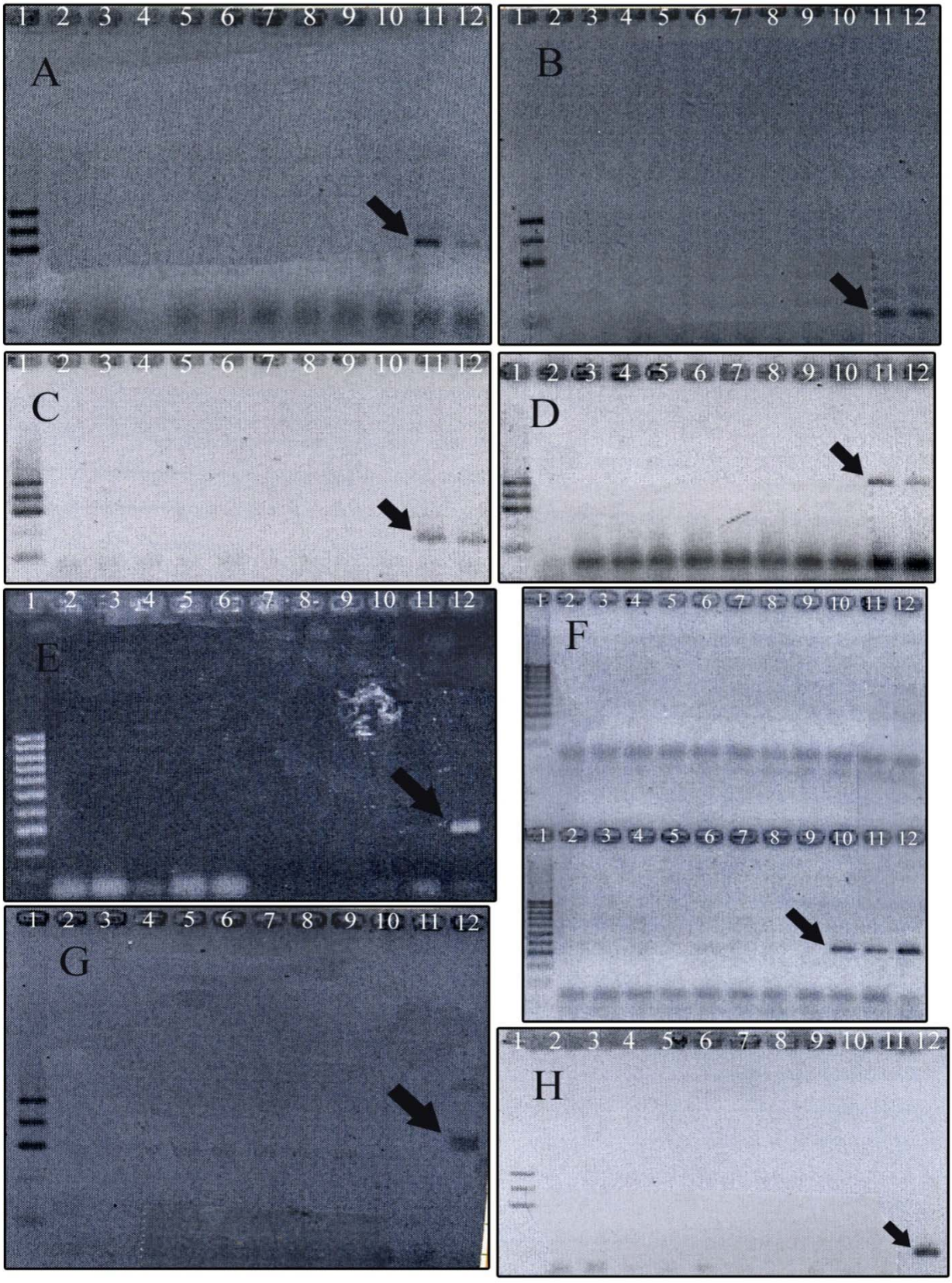
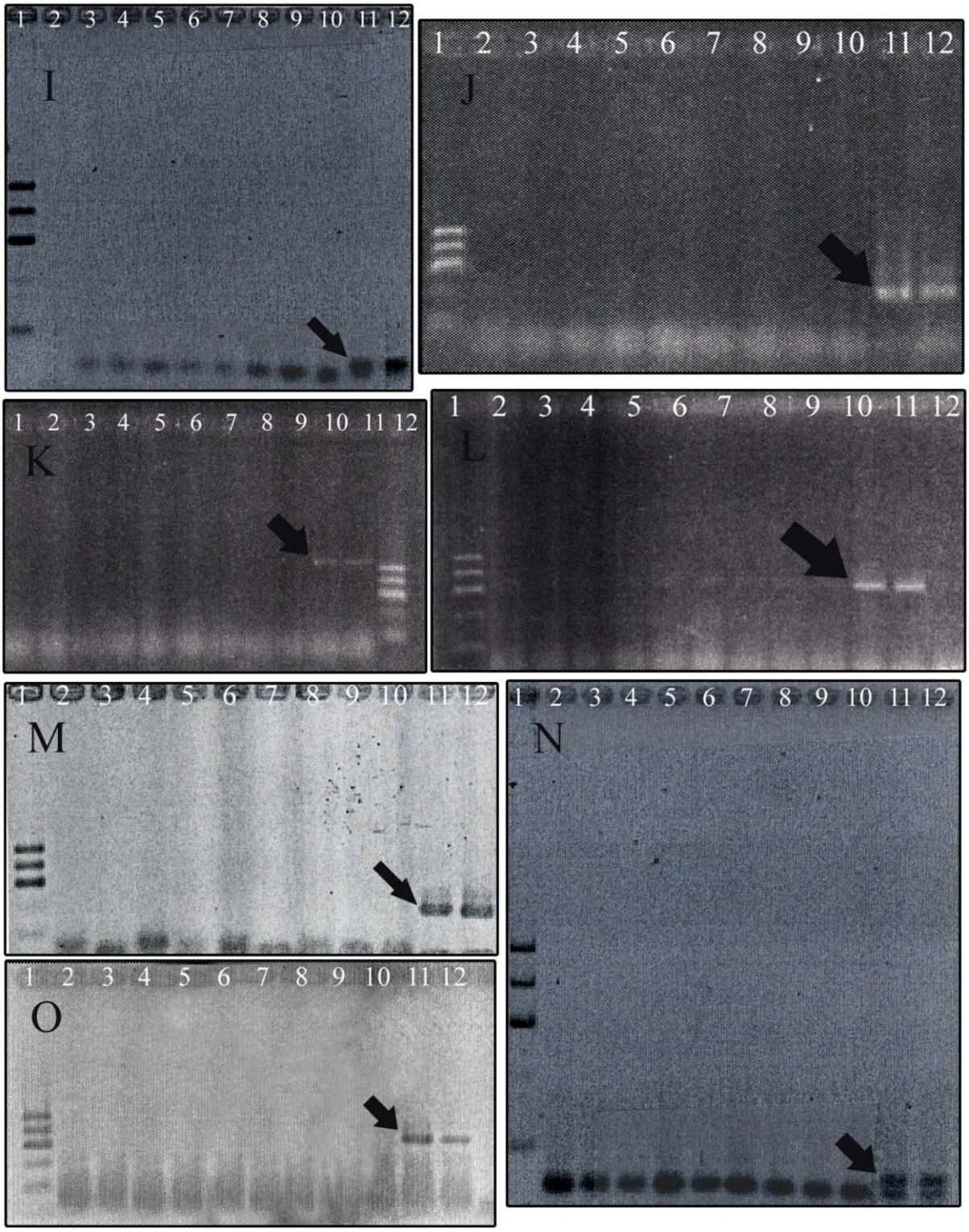


Figure 13 Continued – PCR Amplification of BACs Comprising Minimal Tiling Path Containing *NIP*



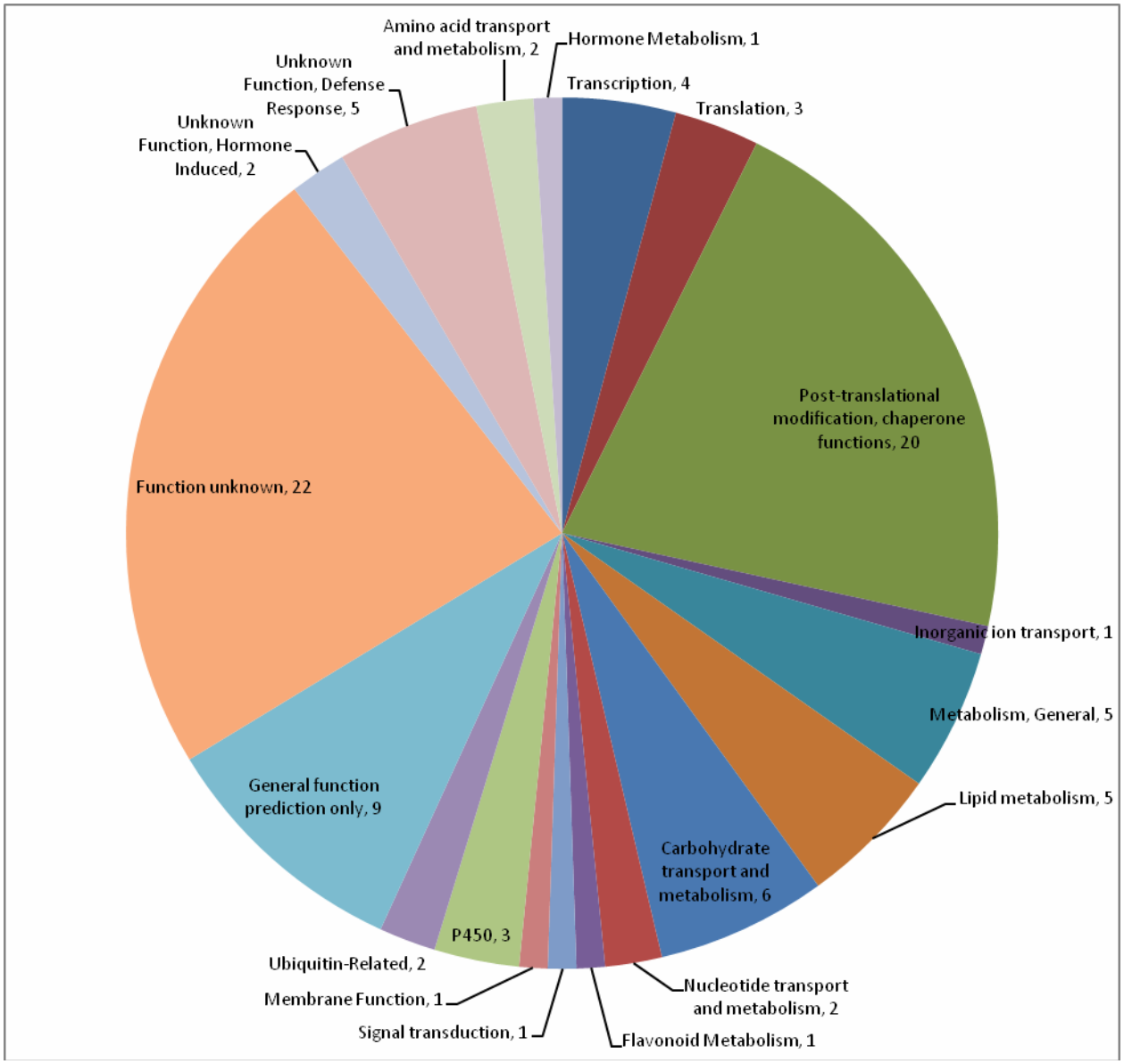
Amplification of candidate transcripts from BACs covering the physical interval containing *NIP*. Arrows on each panel point to the amplified product from positive control

genomic DNA reactions. Each PCR is noted below by its shorthand name (as derived from its predicted function). **A:** SALMSCMT **B:** TC103 **C:** Albumin1 **D:** Lectin Gal1 **E:** THIFOMT **F:** MYB 1 **G:** MYB 2 **H:** GAST **I:** TC102 **J:** HSR201 **K:** ITT **L:** Trypin **M:** TC100 **N:** TC101 **O:** DA4OACT. In Panels **A-D, I, J, M-O**: Lane 1 – 100-500 bp ladder, Lane 2-10 – PCR products with BAC templates - mth2-160M23, mth2-164P8, mth2-187J11, mth2-1G13, mth2-4L4, mth2-71F24, mth2-78L20, mth2-8O7 and mth2-8I23, Lane 11 – A17 Genomic DNA Template, Lane 12 – A20 Genomic DNA Template. In Panels **E, G, H**: Lane 1 – 100-1000 bp ladder, Lane 2-10 – PCR products with BAC templates - mth2-160M23, mth2-164P8, mth2-187J11, mth2-1G13, mth2-4L4, mth2-71F24, mth2-78L20, mth2-8O7 and mth2-8I23, Lane 11 – Water Template (Negative Control), Lane 12 – A17 Genomic DNA Template. In Panel **F**: Lane 1 – 100-1000 bp ladder, Lane 2-12 – mth2-38K19, mth2- 49022, mth2-4L4, mth2-57D22, mth2-68B10, mth2-69F3, mth2-71F24, mth2-78L20, mth2-1G13, mth2-10P15, mth2-J16, Bottom Gel, Lane 1 – 100-1000 bp ladder, Bottom Gel, Lane 2-9 – mth2-144F11, mth2-154G18, mth2-157H22, mth2-160M23, mth2-161P23, mth2-164P8, mth2-187J11, mth2-32AA, Bottom Gel, Lane 10 – A17 Genomic DNA Template, Bottom Gel, Lane 11 – A20 Genomic DNA Template, Bottom Gel, Lane 12 – C90 Genomic DNA Template. In Panel **L:** Lane 1 – 100-500 bp ladder, Lane 2-9 – PCR products with BAC templates - mth2-160M23, mth2-164P8, mth2-187J11, mth2-1G13, mth2-4L4, mth2-71F24, mth2-78L20 and mth2-8O7, Lane 10 – A17 Template, Lane 11 – A20 Template, Lane 12 – PCR products with BAC template - mth2-8I23. In Panel **K**: Lane 1-9 – PCR products with BAC templates - mth2-160M23, mth2-164P8, mth2-187J11, mth2-1G13, mth2-4L4, mth2-71F24, mth2-78L20, mth2-8O7 and mth2-8I23, Lane 10 – A17 Genomic DNA Template, Lane 11 – A20 Genomic DNA Template, Lane 12 – 100-500 bp ladder

2.2.5 Biochemical Function Prediction for Up- And Downregulated Genes

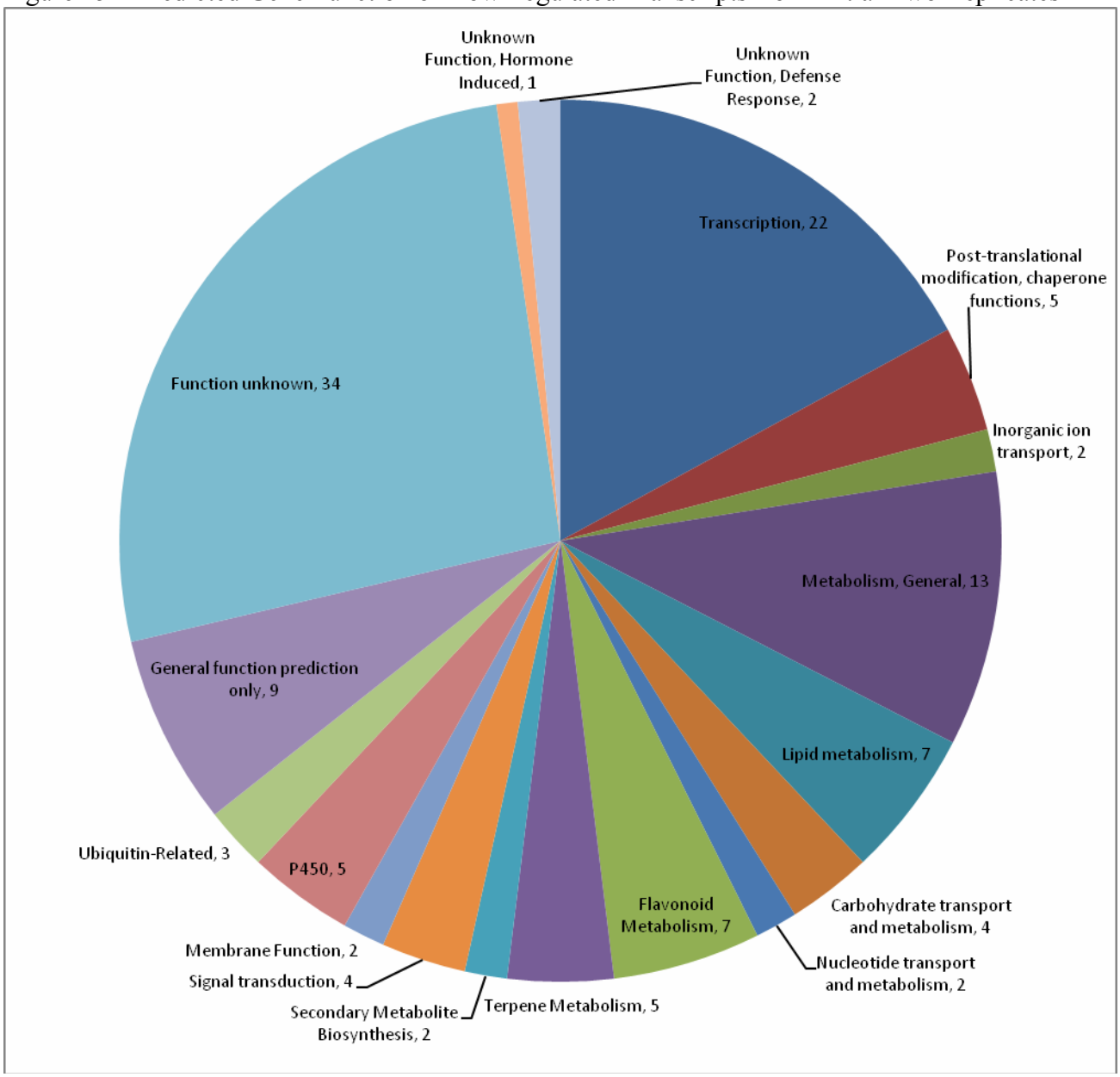
Function analysis was conducted on genes with altered expressions above a minimum average m-value of 2. The genes were classified into general categories based on those found in Clusters of Orthologous Genes (COG) Databases (Tatusov et al., 2001). Figure 14 shows the upregulated genes organized by function and Figure 15 shows downregulated genes similarly organized.

Figure 14 – Predicted Gene Function of Upregulated Transcripts from Initial Two Replicates



Putative functions of gene products that were upregulated in *nip* relative to wild type in the microarray sample data. These predicted functions are based on homology to known *Arabidopsis thaliana*, *Oryza sativa, and Glycine max* genes. Transcripts with unknown function lack sufficient homology to any characterized transcript, while those in the General function category are those with homology to hypothetical gene products in other species.

Figure 15 – Predicted Gene Function of Downregulated Transcripts from Initial Two Replicates



Putative functions of gene products that were downregulated in *nip* relative to wild type in the microarray sample data. These functions are based on homology to known *Arabidopsis thaliana*, *Oryza sativa, and Glycine max* genes. Transcripts with unknown function lack sufficient homology to any characterized transcript, while those in the General function category are those with homology to hypothetical gene products in other species.

2.3 Triplicate Data Screen

The addition of a third data set provided sufficient replicates for further analysis based on both average fold change and standard deviation the combined data. This replicate was extracted, purified and analyzed by microarray by Dr. Catalina Pislariu and Dr. Ivone Torres-Jeves at the Nobel Foundation. The third replicate was analyzed initially using GCOS to directly compare the A17 Replicate #3 to C90 Replicate #3. The aggregate data from all three comparisons using GCOS were combined and the samples were culled in such a way that only transcripts that were differentially expressed in all three replicates remained. The data was then sorted by the average m-value of the three combined replicates and samples below certain average m-values (2 and 1) were removed. The total number of altered expression transcripts identified from this screen are listed in Table 6.

Table 6 – Number of Altered Transcripts Identified by Triplicate Data Screen

Minimum average m-value	Downregulated Transcripts	Upregulated Transcripts
2	70	80
1	370	261

Total number of altered *M. truncatula* transcripts with a minimum average fold change (m-value) between all three replicates.

2.3.1 RMA, Normalization and Chip Quality Assessment

The program RMAExpress was again used as in Section 2.2.2 to calculate a separate set of probe signals based on the original .CEL files of the different replicates. The total data from all six microarrays was analyzed using the RMAExpress program. The output was aggregated in Excel and wild type gene expression was compared to *nip*. The samples were then culled based on the average ratio between A17 and *nip*, with minimum average ratios of 2 and 4. These ratios corresponded to the cutoff points for m-values used previously with GCOS. This data was then

compared to the previous output generated using GCOS (Section 2.3.1). The total number of transcripts with a minimum ratio of 4 (downregulated) or 0.25 (upregulated) can be found in Table 7. Common transcripts identified by both software analyses are in Figures 16 and 17.

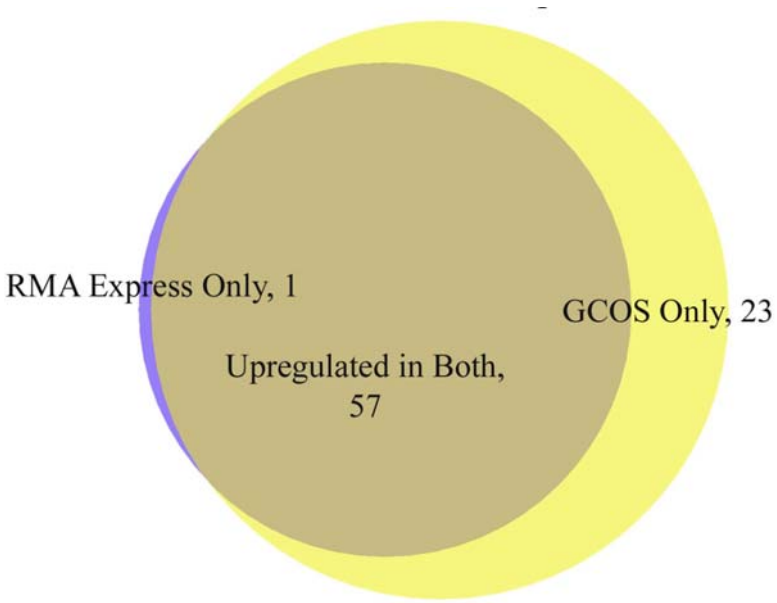
Additional quality assessment for each GeneChip experiment was performed by observing the patterns of altered expression that can be found on each chip in RMAExpress. No macro-scale patterns of distortion were found on these microarrays.

Table 7 – Number of Altered Transcripts Identified In A17 vs. C90 0 DPI Comparison by RMAExpress Software Analysis – Triplicate Data

Minimum Ratio of A17/C90	Downregulated Transcripts	Upregulated Transcripts
4 (Downregulated) 0.25 (Upregulated)	95	58

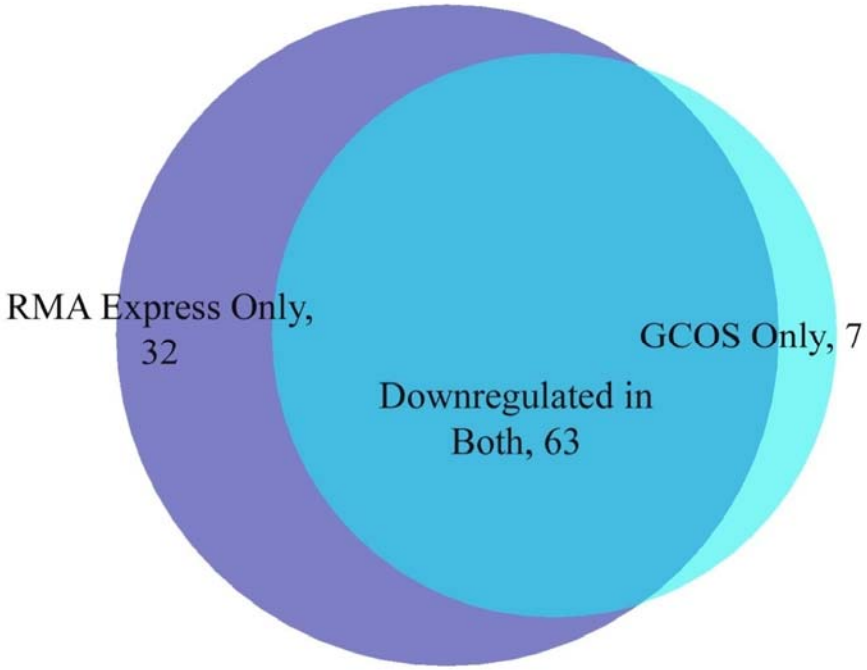
Total number of altered *M. truncatula* transcripts with a minimum average ratio equivalent to an m-value of 2.

Figure 16– Common Upregulated Transcripts between Software Analyses – Triplicate Data



Comparison of upregulated transcripts found by different algorithms above an m-value of 2 (GCOS) or an equivalent ratio of A17 to *nip* of 0.25 (RMAExpress). Of the 58 samples in RMAExpress data set (blue) and the 70 found in the GCOS data set (yellow), 66 of those are the same (beige).

Figure 17 – Common Downregulated Transcripts between Software Analyses – Triplicate Data



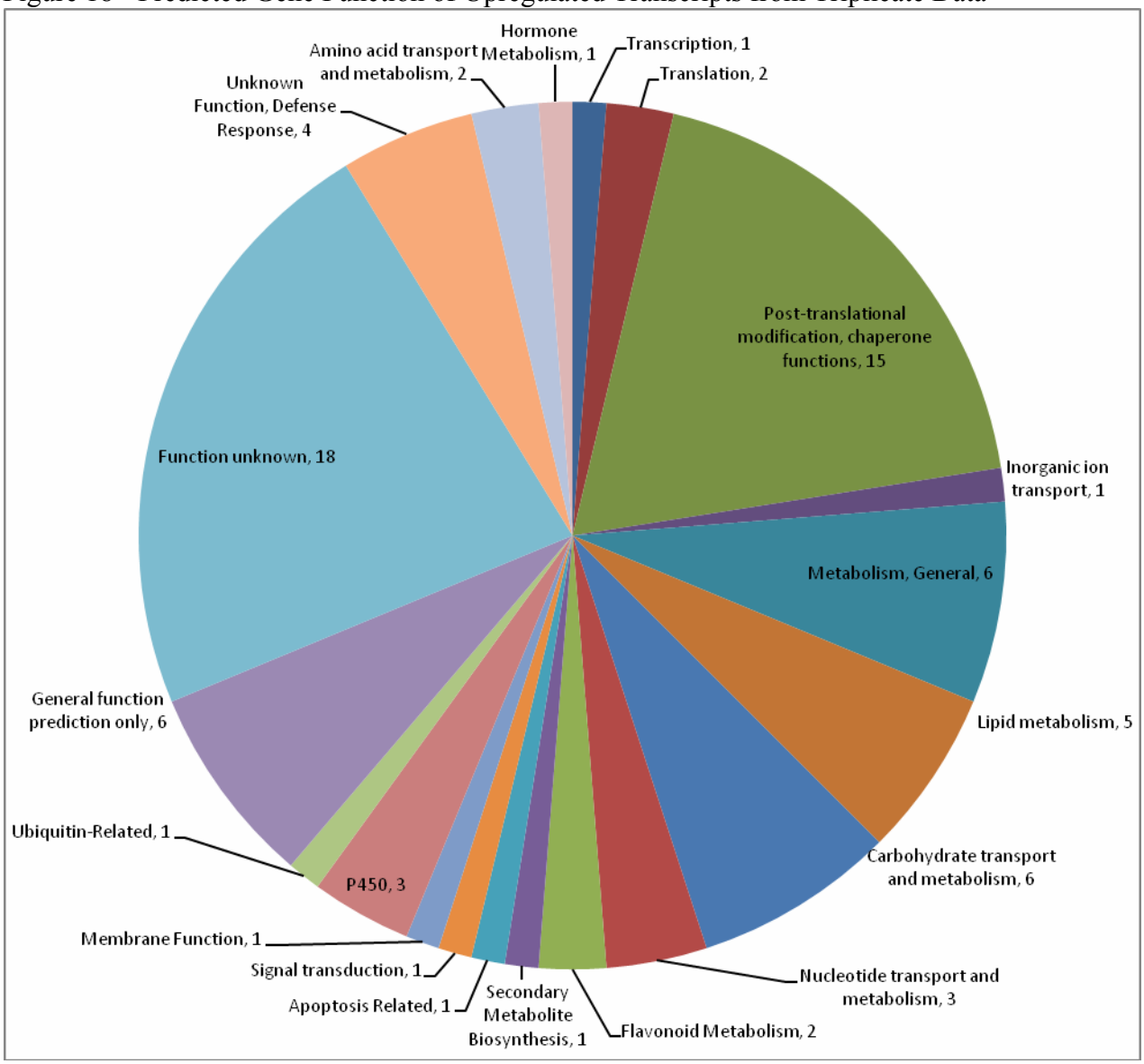
Comparison of downregulated transcripts found by different algorithms below an m-value of -2 (GCOS) or an equivalent ratio of A17 to *nip* of 0.4 (RMAExpress). Of the 95 samples in the RMAExpress data set (purple) and the 70 found in the GCOS data set (light blue), 63 of those are the same (blue).

2.3.2 Biochemical Function Prediction of Up- And Downregulated Triplicate Transcripts

Function analysis was conducted on genes with altered expressions above a minimum average m-value of 2. Figures 18 (upregulated genes) and 19 (downregulated genes) represent these different gene functions in a graphical format. Common transcripts between the two data sets that were generated from the data are shown in Figures 20 (upregulated) and 21 (downregulated).

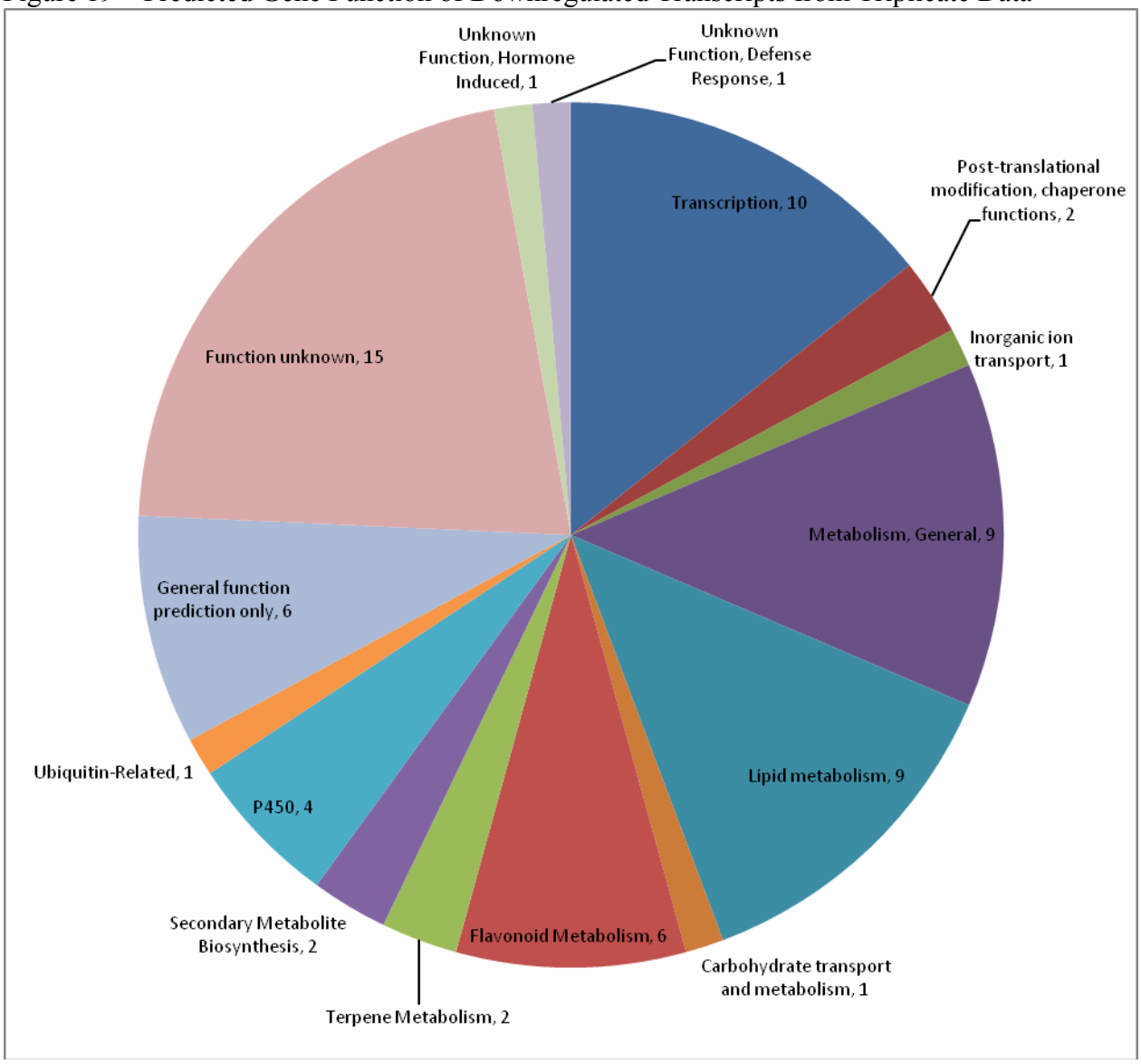
Specific changes in flavonoid metabolism were found in both upregulated and downregulated transcripts. The major pathway of flavonoid metabolism and corresponding changes in predicted enzymes of that pathway are in Figure 22.

Figure 18 - Predicted Gene Function of Upregulated Transcripts from Triplicate Data



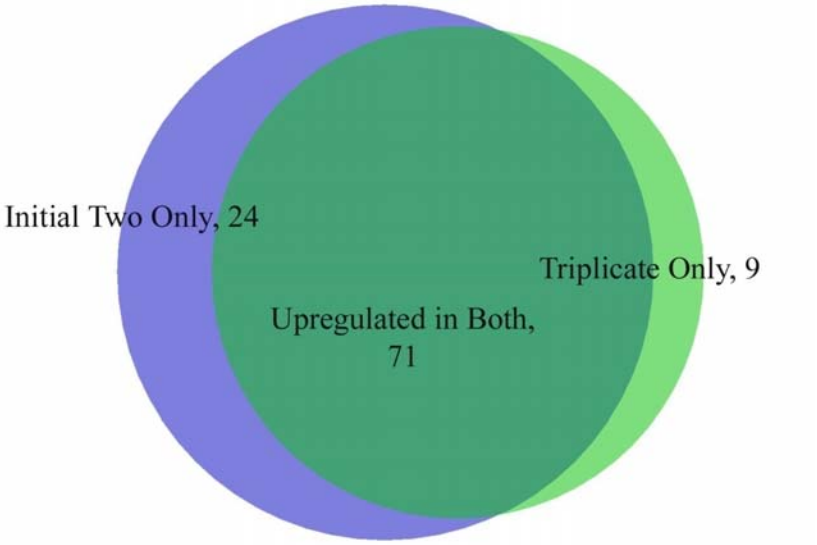
Putative functions of gene products that were upregulated in *nip* relative to wild type in the microarray sample data. These functions are based on homology to known *Arabidopsis thaliana*, *Oryza sativa, and Glycine max* genes. Transcripts with unknown function lack sufficient homology to any characterized transcript, while those in the General function category are those with homology to hypothetical gene products in other species.

Figure 19 – Predicted Gene Function of Downregulated Transcripts from Triplicate Data



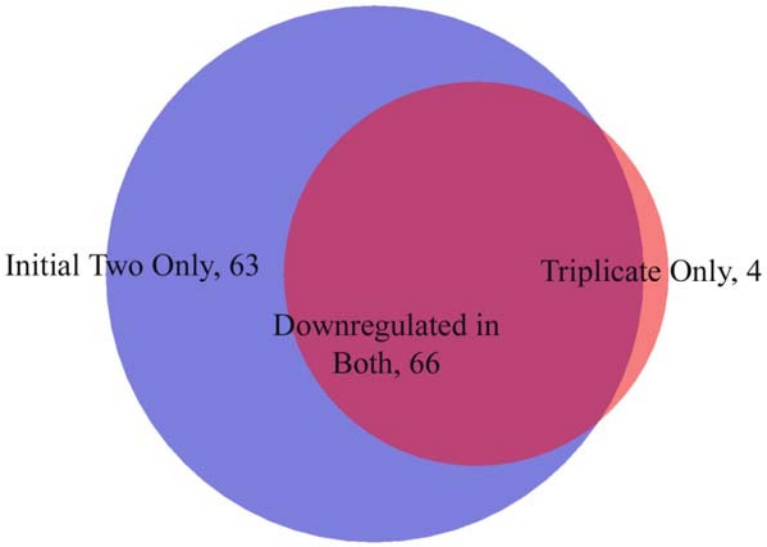
Putative functions of gene products that were downregulated in *nip* as compared to wild type in the microarray sample data. These functions are based on homology to known *Arabidopsis thaliana*, *Oryza sativa, and Glycine max* genes. Transcripts with unknown function lack sufficient homology to any characterized transcript, while those in the General function category are those with homology to hypothetical gene products in other species.

Figure 20 – Common Upregulated Transcripts between Initial Two Replicates and Triplicate Datasets



Comparison of upregulated transcripts above an m-value of 2 in duplicate vs. triplicate data analysis. Of the 95 samples in the initial data set (light blue) and the 80 found in the triplicate data set (light green), 71 of those are in common between sample sets (dark green).

Figure 21 – Common Downregulated Transcripts between Initial Two Replicates and Triplicate Datasets



Comparison of downregulated transcripts below an m-value of -2 in duplicate vs. triplicate data analysis. Of the 129 samples in the initial data set (light blue) and the 70 found in the triplicate data set (pink), 66 of those are the same (purple).

2.4 A17 as. C90 - 10 DPI Comparison Data

In order to discern temporal gene expression patterns in *nip,* an additional microarray experiment was conducted at the 10 dpi time point in nodulation, or ten days post inoculation with rhizobia. At this time point in wild type nodule development, the nodules are well formed physically and beginning to fix nitrogen, whereas *nip* nodules are small and have begun to accumulate brown phenolic pigments. By analyzing the expression patterns of *nip* at this time point, we can determine which genes found to be differentially regulated in the 0-day analysis data are also symbiotically regulated. A single microarray replicate of A17 vs. C90 at 10 DPI was conducted by Dr. Catalina Pislariu at the Noble Foundation. This data was analyzed using GCOS to compare the wild type expression levels to those of *nip.* Without replicate data, removal of the large numbers of false positive data points that would be present in the sample is not possible, and thus the data was used mainly as a reference point for genes identified in the 0-day screen. Table 8 shows the total number of altered transcripts identified above a minimum fold change.

Of the transcripts identified by the 10-dpi microarray screen, 44 upregulated transcripts and 22 downregulated transcripts experience similar expression patterns at 0-dpi. Notable in this data is the absence of all but 1 of the previously identified heat shock proteins are upregulated at 10 dpi.

Table 8 – Number of Altered Transcripts Identified In A17 vs. C90 10 DPI Comparison

Minimum m-value	Downregulated Transcripts	Upregulated Transcripts
2	1012	726
1	1804	1863

Total number of altered *M. truncatula* transcripts with a minimum fold change (m-value) between wild-type A17 and *nip*.

CHAPTER 3

DISCUSSION

3.1 GeneChip Quality Control Assessment

Multiple experimental controls were performed during the course of work to assess the quality of the generated data. The RNA samples selected for the microarray experiment were determined to have the highest quality (Table 1) of the multiple technical replicates generated. This ensured that the data from these samples would best represent the *nip* transcriptome compared to wild type. Real-time RT-PCR controls (Table 5) confirmed that the expression levels and up- or down expression calls in *nip* compared to wild type made by the GCOS software were accurate. Real-time RT-PCR provides a more accurate measure of transcript abundance as compared to microarray quantification because of the ability of real-time RT-PCR to effectively measure very low levels of transcript which is difficult in microarray readings.

Addition of a third replicate enabled us to decrease the number of presumably false positive expression signatures (Figures 20 and 21) with a corresponding increase in the accuracy of the dataset. Furthermore, comparisons of the data using different algorithms from separate analysis software programs yielded very similar outputs of gene expression in the *nip* mutant (Figures 8, 9, 16 and 17). Finally, analysis of macro scale trends of expression on each chip (Figure 10) showed that all of the chips had very little large scale distortion with the exception of some negative values on the *nip* Day 0 Replicate 1 chip. These analyses show that the experiments are sufficiently accurate in their representation of the *nip* transcriptome.

3.2 Analysis of Downregulated Transcripts For Candidate Genes

Attempts were made to identify the *NIP* gene from transcripts that were downregulated in *nip* relative to wild type. Such an approach has been shown to work previously in *Medicago* and in barley. Because of absence of a transcript in deletion mutations or preferential degradation of a mutated transcript in point mutations, the mutant transcript may be less abundant, which can be detected in microarray analysis (Mitra et al., 2004). For *nip,* which is an EMS-generated mutant, this approach may or may not be applicable depending on the level of alteration to the transcript and subsequent degradation of the mRNA. Neither the bioinformatic nor the PCR amplification analyses (Figure 12) of downregulated transcripts in the *nip* mutant were unable to identify any lowered expression transcripts in the physical interval containing the *NIP* gene (Section 1.8). From the original set of 129 highly downregulated transcripts, 31 were eliminated based on sequence homology and physical location in the *M. truncatula* genome. The addition of the triplicate data set removed 63 transcripts from our search and of those remaining, 20 were eliminated based on low signal strength, predicted biochemical function or false-positive data as determined by inconsistency between replicates. The remaining 15 transcripts were selected for testing by PCR amplification from the physical interval containing *NIP*. These were tested by PCR amplification and were found not to be contained in the BACs comprising the minimal tiling path that contains *NIP* (Figure 12).

Based on this data, I conclude that it is unlikely that the *NIP* gene is preferentially degraded in the *nip* mutant. Therefore it is not identifiable based on expression data because as an EMS generated mutant, there may not be sufficient alteration of the mRNA for the cellular machinery to target the *nip* transcript for degradation. The original work on which the clone-by-transcript concept in *Medicago* is based (Mitra et al., 2004) utilized a deletion mutant, *DMI3,*

that was generated by fast neutron bombardment. Additionally, Mitra et al. showed that the previously characterized transcript for the EMS-generated mutant *DMI2* could be identified using transcriptional data, and that similar results could be obtaining using a point mutant of barley (2004). However, the degradation of the transcript in a point mutant would be contingent on there being sufficient alteration of the transcript such as an additional stop codon, and thus the technique is not be applicable in the case of the *nip* mutant.

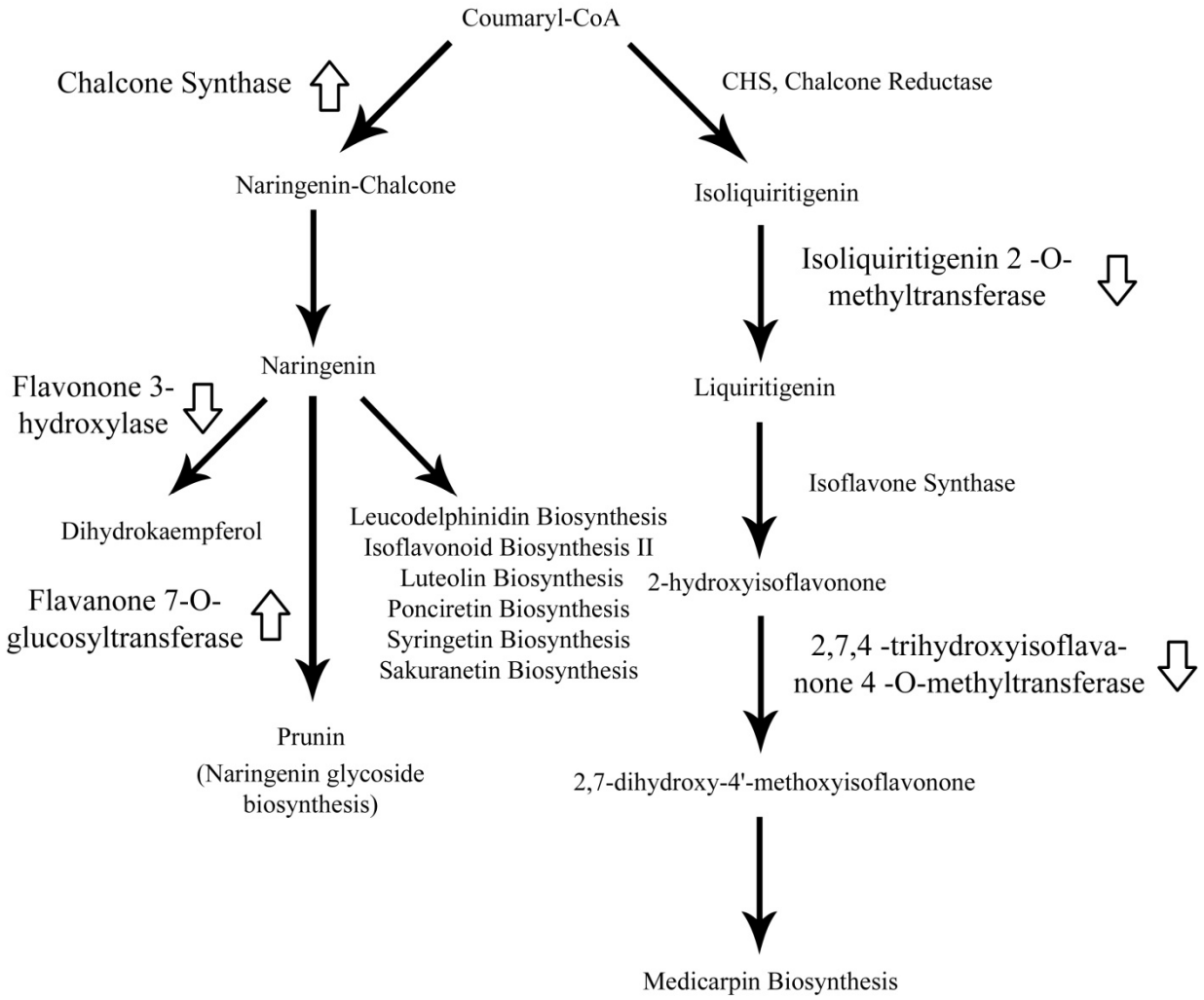
3.3 Functional Analysis of Gene Transcripts

Putative gene product functions were assigned to each altered transcript (Figures 14, 15, 18 and 19) identified in the 0-Day expression profiles. This data showed that the 0-DPI *nip* plants have increased expression of stress related genes. The most notable of these gene groups is the large number of upregulated heat shock proteins (HSPs) found in the mutant transcriptome, such as Hsp22.7, Hsp70, Hsp17, Hsp17.9, Hsp20, Hsp17.7, Hsp18.2 and Hsp22. This is consistent with previous work by Veershlingam et al. showing increases in stress after nodulation (Veershlingam et al., 2004). This new data suggests that the *nip* mutant experiences stress prior to becoming infected by rhizobia and the basal defect in *nip* is a heightened stress response. One possibility is that *NIP* encodes as suppressor of basal resistance or stress responses and the *nip* mutation abolishes the function of such a gene. However, analysis of these HSPs in the 10-dpi microarray data has shown that all but 1 HSP are absent from the list of upregulated transcripts. While other HSPs are upregulated in the 10-dpi data, this trend is suggestive of changes in stress upon inoculation with rhizobia and subsequent nodulation.

Functional analysis has also shown alterations in flavonoid metabolism, with genes involved in chalcone and naringenin glycoside biosynthesis being upregulated while those

involved in medicarpin and isoflavonoid pathways are downregulated. The transcripts for 2,7,4 -trihydroxyisoflavanone 4 -O-methyltransferase, isoliquiritigenin 2 -O-methyltransferase and flavanone 3-hydroxylase are downregulated, while those for UDP-glycose:flavonoid glycosyltransferase and chalcone synthase (CHS) are upregulated. The effect of these changes on flavonoid metabolism (Figure 22) could lead to decreased medicarpin biosynthesis and increased naringenin and naringenin glycoside biosynthesis. This may play a role in the altered symbiotic phenotype of *nip* due to the disruption of nodulation precursor signals being exuded by the plant into the rhizosphere. Additionally, recent work has shown that flavonoids have an inhibitory effect on polar auxin transport in *Medicago* as well as in *Arabidopsis* (Wasson et al., 2006; Zhang et al., 2007). Wasson et al. also showed that flavonoids are required for nodulation and implied that the inhibitory effect of flavonoids on polar auxin transport might be necessary during nodulation (2006). Within the context of our work, this suggests that possibly auxin transport is disrupted in *nip* due to the alterations in flavonoid metabolism. Furthermore, analysis of the 10-dpi *nip* transcriptome shows that alterations in flavonoid metabolism are present in *nip* after inoculation with rhizobia. At 10-dpi, expression of CHS is highly disrupted as evidenced by both highly upregulated and highly downregulated transcripts in 10-dpi *nip*. Transcripts with predicted functions involving isoflavonoid metabolism are also highly downregulated and upregulated in *nip* at 10-dpi. These disruptions in flavonoid metabolism suggest that the *nip* mutant continues to have altered flavonoid metabolism after inoculation and may have corresponding changes in polar auxin transport as well. Disrupted auxin transport affecting nodulation could potentially be the underlying cause of the *nip* nodule phenotype and/or the lateral root phenotype.

Figure 22 – Pathways of Flavonoid Metabolism Altered in Triplicate 0-Day *nip*



Alterations in flavonoid metabolism as predicted by putative transcript function of transcripts with highly altered expression (m-value is greater than 2). Hollow arrows indicate the change in expression of the corresponding adjacent enzyme, based on predicted function. Black arrows represent biochemical reaction steps.

3.4 Establishment of an Expression Profile Database For *nip*

A database containing all expression data generated for both 0 dpi and 10 dpi comparison experiments was generated for the purpose of providing an easily searchable reference for the screening of candidate genes. The physical and genetic mapping of the *nip* gene has narrowed down a region of LG1 that is currently being sequenced by Dr. Bruce Roe's research group.

With the full sequence of the area, it should be possible to screen for candidate genes by comparing the genomic sequence to transcript sequences probed by the microarray. Furthermore, gene prediction from the genomic sequence could suggest candidate genes based on phenotype, which can then be compared against microarray transcripts. Using this data, candidate genes can be quickly assessed for their presence or absence (along with actual signal data) in the A17/*nip* comparison data, such that a candidate gene that has no measurable expression in both wild type and mutant samples can be assumed to not be *nip.* This methodology is currently being employed by Dr. Rebecca Dickstein with the assistance of Yi-Ching Lee in screening of candidate genes.

CHAPTER 4

METHODS

4.1 Aeroponic Chamber Growth and Harvest of Plants

For the 0-Day C90 and A17 roots, germinated seedlings were grown in aeroponic chambers on nitrogen-free, full-strength Lullien media for 5 days at 22° C in 16 hours light/8 hours dark and then harvested to provide the 0-day time point before bacterial interaction begins (Lullien et al., 1987). Plant roots were harvested from the aeroponic chamber by slicing them at the base of the shoot, with care to not harvest the green shoot material. Plant material was flash-frozen in liquid nitrogen and stored at -80° C.

Two biological replicates (two separate aeroponic chambers) of A17 (wt) and C90 (*nip*) were grown at different times and harvested similarly. RNA from each of these replicates was extracted for microarray analysis.

The third replicate was grown under identical conditions but additionally contained plants for harvest at later time points. 10-dpi plants were used in a separate microarray comparison.

4.1.1 Seed Germination and Chamber Loading

Seed pods were put into a seed mill and individual seeds were recovered to a 50 ml conical tube to which 5 volumes of concentrated sulfuric acid were added for 5-10 minutes while being gently agitated continually. The acid was decanted the seeds were and washed 5 times with sterile ddH_2O. The seeds were then surface sterilized using 5 volumes of commercial grade bleach (~5% sodium hypochlorite) for 3 minutes. The bleach was decanted and seeds were rinsed 7-8 more times in sterile ddH_2O. A total of 10 successive sterile ddH_2O washes for 30 minutes each served to imbibe the seeds. Seeds were placed in fresh sterile water and left at 4° C overnight. The next day, two more rinses of sterile ddH_2O were used to remove any seed

exudates. The seeds were plated on 100x15mm petri dishes containing a thin layer of germination media with a disc of sterile filter paper on top. Germination media consists of 1 mM magnesium sulfate, 0.75 mM potassium phosphate (KH_2PO_4), 1 mM sodium phosphate (Na_2PO_4), 15 μM ferric citrate, .75 mM calcium nitrate, 0.7 mM calcium chloride, 0.35 μM cupric sulfate, 4.69 μM manganese sulfate, 8.46 μM zinc sulfate, 51.3 μM boric acid, 4.11 μM sodium molybate and 0.9% agar. Seeds were spread across the surface of the filter paper with care not to have seeds touching. The seeds were then placed upside down in a dark drawer overnight at room temperature to allow for germination. The following day germinated seedlings were loaded into an aeroponic chamber containing 1X Lullien media. Seedlings are placed in a hole in plastic wrap on the lid of the chamber so that the roots are hang down through the lid into the aeroponic chamber.

4.1.2 Aeroponic Chamber Preparation

We use a protocol modified from the Cook Laboratory (Cook, 2001). Prior to seed treatment, an aeroponic chamber was suitably prepared for the plants. The aeroponic chamber system used consisted of a motor apparatus (a modified Defensor 505 humidifier) in a 30 gallon garbage can. The entire system was sterilized by bleach for a minimum of 30 minutes and washed a minimum of 5 times with de-ionized water to prevent contamination with *S. meliloti*. The chamber was then filled with 8 liters of sterile Lullien media. Lullien media contains 0.52 mM K_2SO_4, 0.25 mM MgSO4, 1 mM $CaCl_2$, 50 μM Na_2 EDTA, 30 μM H_3BO_3, 10 μM $MnSO_4$, 0.7 μM $ZnSO_4$, 0.2 μM $CuSO_4$, 1 μM Na_2MoO_4, 0.04 μM $CoCl_2$, 5 mM NH_4NO_3, 5 mM K_2HPO_4, 5 mM KH_2PO_4, 33.3 μM Fe(II)Cl_2 and 23.8 μM Fe(II)SO_4. Full strength (1X) media containing no nitrate (NH_4NO_3) was used.

4.1.3 Inoculation with Rhizobia

The aeroponic chamber for the third replicate was inoculated with a culture of *S. meliloti* after 5 days of growth on nitrate-free media. For inoculation, a 50 ml culture of bacteria in TY media was decanted directly into the chamber. The *S. meliloti* strain used was Rm1021 pXLDG4 that carries a LacZ reporter gene for staining of the bacteria in roots (Boivin et al., 1990).

4.1.4 Growth Conditions

Medicago plants in a growth room at 22° C. Day length was maintained at 16 hours light/8 hours of darkness.

4.2 RNA Extraction and Preparation

For RNA extraction, the protocol of Dunn et al was followed (Dunn et al., 1988). Briefly, RNA from roots was extracted by manual disruption of frozen tissue under liquid nitrogen with a chilled mortar and pestle. After manual disruption of the tissue, the tissue powder was then moved into a tube of DEPC-treated RNA extraction buffer consisting of 0.2 M sodium acetate (pH 5.5), 1% SDS, 10 mM EDTA (pH 8). The buffer was mixed with a volume of 4:1 phenol:chloroform (chloroform mixture consists of 24:1 chloroform:isoamyl alcohol) and the liquid phase was extracted. A second identical phenol:chloroform extraction was then conducted, followed by a final extraction in one volume of chloroform. Samples were then precipitated using 10 M LiCl, frozen at -80° C for 15 minutes and centrifuged at 11,000 g for 15 minutes to form an RNA pellet. The RNA was resuspended then reprecipitated with 1/10 volume of 10 M LiCl and 2.5 volumes of 100% EtOH, frozen at -80° C for 15 minutes and

centrifuged again at 11,000 g for 15 minutes. The supernatant was then removed and the RNA pellets were resuspended in 40 µl of DEPC-treated ddH_2O. A total of 3 identical tube sets were generated from each batch of ground tissue, which were then recombined into a single 120 µl sample for DNase digestion and sample purification.

A modified QIAGEN® RNeasy® protocol was used to degrade genomic DNA in the sample and remove contaminants from the sample (QIAGEN Inc., Hilden, Germany, http://www1.qiagen.com/). This protocol was provided by the González Lab (Juan González, Natalie Gurich, and Jennifer Morris) at the University of Texas at Dallas and was originally created by Elizaveta Krol and Eva Schulte-Berndt (University of Bielefeld) as a modified QIAGEN protocol. Briefly, a QIAGEN On-Column DNase digestion was performed using the RNeasy Mini-Spin column using the standard QIAGEN protocol, followed by a second digestion with Ambion® DNase using the standard Ambion reaction conditions (Ambion Inc., Austin, TX, http://www.ambion.com/). A second RNeasy Mini-Spin Column was used to remove nucleases and the final product was eluted from the column in RNAse free water. Each sample was then separated into aliquots and for quality controls and frozen at -80° C as follows: 2 µl for spectrophotometer quantification, 2 µl for agarose gel qualification, 2 µl for Agilent Bioanalyzer 2100 qualification, 2 µl as a backup sample, 12 µl for genomic real-time PCR control quantification, and 15 µl for reverse transcription real time PCR control quantification. The rest of sample remained (unthawed) in original tube for final Affymetrix® application (Affymetrix Corp., Santa Clara, CA, http://www.affymetrix.com).

4.3 Determination of RNA Integrity for Microarray Applications

To ensure that the extracted RNA was sufficiently intact for the purpose of microarray analysis, three quality control checks were performed on each sample. Initially, 2 μg total RNA was converted into cDNA using the New England Biolabs protocol for cDNA preparation with a gene specific primer set (New England Biolabs, Beverly, MA). A primer set that amplified part of the gene Msc27 was used, because it is constitutively expressed but has but a single copy in the genome (Allison et al., 1993; Csanadi et al., 1994; Dickstein et al., 2002). Real-time polymerase chain reaction (Real Time PCR) was used to ascertain the relative abundance of the transcript. Real Time PCR of the Msc27 gene was performed using the Cephied Smart Cycler 2, using a mix of 2.5 μl of each 3 mM oligonucleotide primer (F and R), 2.5 μl of 5x SYBR Green (diluted in sterile water from 10,000x SYBR Green/DMSO stock), 1-5 μl of cDNA from a 0.08 μg/μl reverse transcription reaction and an appropriate amount of water to bring the sample volume to 25 μl. Takara OmniMix PCR beads, containing the polymerase, dNTP's and appropriate buffers and cofactors, were used in the reaction mixture, with 1 bead supplying the necessary reagents for two 25 μl reactions. 40 reaction cycles of 60 C for 40 seconds followed by 94 C for 20 seconds were conducted. SYBR Green fluorescence was monitored by fluorescence excitation/emission at 498/522 nm respectively during the 60 C cycle. A melt curve determining the binding of SYBR Green to DNA strands was performed to determine the dissociation temperature of the products.

To determine whether any residual genomic DNA remained in the samples, an identical Real Time PCR reaction mixture was used on each sample, with the exception that 1 microgram of untreated RNA used as the template rather than 5 μl of cDNA.

Final quality control analysis of the RNA samples was conducted using the Agilent 2100 Bioanalyzer and the RNA 6000 Nano Lab-on-a-Chip kit. Each 1 µl sample was placed into individual wells in the microfluidics chip, according to the Agilent protocol (Agilent, 2006). Samples were analyzed by the Agilent software and were assigned a RNA Integrity Number (RIN) based on the relative background (degradation) of each sample and the ratio of 28S to 18S ribosomal RNA (Meuller et al., 2004). Samples selected for microarray analysis were selected for the highest RIN rating (on a 1-10 scale), with a preferred number being greater than 9.0.

4.4 Affymetrix Microarray Experiment

The Affymetrix GeneChip® hybridization for the first two replicates was carried out in the manner described in the GeneChip Expression Analysis Manual by technicians at the University of Texas Southwestern (Affymetrix, 2004). Each of the 4 samples was run on a separate GeneChip; 2 wild type A17 samples from different aeroponic chambers and their corresponding *nip* mutant samples. Data from the experiment was provided in the form of Gene Chip Operating System (GCOS) readable files.

The third dataset replicate for the *nip* 0-day comparison and the initial dataset from the *nip* 10-day comparison were conducted at the Nobel Foundation by Dr. Catalina Pislariu and Dr. Ivone Torres-Jeves and the initial RMAExpress analysis of that data was generated by Dr. Yuhong Tang of the Noble Foundation (Bolstad, 2007).

4.5 Expression Analysis in GCOS

Data from each chip was analyzed using the GCOS software, with relative expression levels for each probe set calculated using each chip's internal controls, along with p-values

indicating the relative precision of each measurement. Biological samples were compared to one another (A17-1 vs. C90-1, A17-2 vs. C90-2) and fold changes between wild type and mutant were calculated. This fold change data was exported from GCOS into an XLS spreadsheet for use with Microsoft Excel. Using Excel, the two data sets were combined and then sorted based on which transcripts were differentially regulated (Increase/I or Decrease/D) in order to determine which probes had similar expression changes between the replicates. Transcripts that were differentially regulated in both replicates were then sorted by their average m-value (Log_2 fold change).

Analysis of the triplicate data was performed in a similar fashion but with the further removal of transcripts based on their expression call in the third replicate.

4.6 Transcript Screening for Candidate Genes

Probe sets were then analyzed using the data available from the *Medicago* Sequencing Project (medicago.org), the Dana Farber Cancer Institute (DFCI, formerly The Institute for Genomic Research/TIGR) and the National Center for Biotechnology Information's (NCBI) BLAST database. Probes were analyzed to predict function and also to attempt to identify a physical location of the probe in the *Medicago* genome.

To further characterize the data from the microarray output, the most highly upregulated and downregulated transcripts were analyzed based on currently available genomic data from medicago.org, DFCI and the NCBI. Each TC or BAC predicted gene fragment from the chip was checked using the DFCI database for new annotations that were made since the creation of the Medicago Library File for the GCOS software suite using the *Medicago truncatula* Gene Index (MTGI) 7.0, which has since been updated to MTGI 8.0. Each probe set produced from

BAC sequence was used to search against the medicago.org genome project to identify any BACs located at the top of chromosome 1, the known location of the *nip* gene. By process of exclusion, any BAC probe set in the downregulated list that was shown to conclusively localize to a different chromosome was excluded from further molecular biology screening assays for possible *NIP* transcripts. For probes created from tentative consensus sequences, the entire sequence from the TIGR database was used to determine whether the TC had distinct homology to a given published *M. truncatula* BAC clone in the NCBI database. If the homology was above 95-97% in contiguous regions of the transcript, we could then say that the transcript is likely to be located on that particular BAC. The location of the BAC could then be identified using the medicago.org database and the transcript could be localized to a single chromosome. With additional BAC clone sequence being added to the database daily as part of current efforts to sequence *M. truncatula*, this data was revisited to ensure that the BLAST searches were compared to all available data before molecular test approaches were conducted for the localization of possible *NIP* transcripts.

4.7 Predicted Biochemical Function Analysis of Transcripts

Highly up- or downregulated (>2 m-value) transcripts were subjected to functional analysis prediction. The highest homology of each transcript was is noted by the description category associated with each transcript in GCOS. Based on this homology, predicted gene functions were assigned for each transcript based on data regarding in the following databases: MetaCyc (www.metacyc.org) , BRENDA (http://www.brenda.uni-koeln.de), ExPASy (http://expasy.org) and KEGG (http://www.genome.jp/kegg).

This was then used to a category based on those found in the Clusters of Orthologous Groups (COG) database (Tatusov et al., 2001). Additional categories were assigned as needed based on symbiosis-related gene products, such as flavonoid metabolism and genes of unknown function that are elicited in defense responses.

4.8 Microarray Confirmation by RMAExpress Analysis

Confirmation of data created using GCOS was conducted by using a secondary program, RMAExpress (Bolstad, 2007). RMAExpress was used for normalization and Robust Multichip Average (RMA) analysis. This program compares internal values of GeneChips for normalization but has no inherent functionality for comparison of separate chips to one another. The original data files from GCOS (.CEL) were analyzed by the program and output as spreadsheets containing the signal values for each transcript. Microsoft Excel was then used to create a ratio of the A17 wild-type signal to the mutant *nip* signal for each replicate. The average ratios between replicates were then used to sort the data into lists for analysis.

4.9 Transcript Probing of the Genetic Interval

4.9.1 Extraction of Genomic DNA

Genomic DNA from the *M. truncatula* ecotypes A17 and A20 were used as positive controls for BAC amplification reactions. Genomic DNA samples were extracted from a single leaf of each plant which was flash frozen in liquid nitrogen in a 1.5 ml eppendorf tube and ground into a powder using a plastic pestle. 700 µl of 65°C CTAB reagent was added to the tissue powder. CTAB reagent consists of 2% w/v Cetyl Trimethyl Ammonium Bromide (CTAB), 100 mM Tris, 20 mM EDTA, 1.4M NaCl, 2% Polyvinylpyrrolidone (PVP) and 0.2%

v/v of Beta-mercaptoethanol (BME) added just before use. Samples were incubated at 65°C for 15-30 minutes. 570 µl of 24:1 chloroform:isoamyl alcohol was added to each sample. Samples were mixed by light inversion and then centrifuged for 5 minutes at 11,000 x g. The aqueous layer was removed to a second eppendorf tube and 0.7 volumes of isopropanol was added and mixed by inversion. The samples were spun at 11,000 x g for 10 minutes. The supernatant was removed and 500 ml of cold 70% ethanol was added. The samples were centrifuged again at 11,000 x g for 5 minutes. The supernatant was removed and the pellets were allowed to dry, at which point they were suspended in 20 µl of sterile water. 10x dilutions of this genomic DNA was used as template in PCR amplification.

4.9.2 PCR of BACs on Genetic Interval

Primer sets were made for specific highly downregulated genes, which were then probed against different BACs representing the putative area of chromosome 1 where the *NIP* gene is located. PCR was used to determine if any probe sets localized these BACs. As controls, genomic DNA was used as template for each PCR reaction. Each sample of 20 µl was mixed using the following reagents:

1.6 µl - 2.5 µM dNTPs

1.5 µl - 5 µM forward primer

1.5 µl - 5 µM reverse primer

2 µl - 10x Thermopol buffer (New England Biolabs) - 20 mM Tris-HCl (pH 8.8, @ 25°C), 10 mM KCI,10 mM $(NH_4)_2SO_4$, 2 mM $MgSO_4$, 0.1% Triton X-100

0.1 µl (0.5 units) - Taq DNA Polymerase (New England Biolabs)

12.3 –sterile ddH_2O.

For template, 1 µl of 100x diluted BAC extract was used per each 20 µl reaction. 1 µl of genomic DNA from A17 and A20 ecotypes respectively was used as template for the positive control reaction. 1 µl of sterile ddH_2O was used as a negative control.

Reaction conditions were as follows, unless noted:

Step 1: 94°C - 5 minutes
Step 2: 94°C - 30 seconds
Step 3: 56°C - 30 seconds
Step 4: 72°C - 30 seconds
Repeat Steps 2-4 for 35 total cycles
Step 4: 72°C – 5 minutes
Step 5: 4°C – Hold indefinitely

The BACs used, which cover the genetic interval are mth2-160M23, mth2-164P8, mth2-187J11, mte1-1G13, mth2-4L4, mth2-71F24, mth2-78L20, mth2-8O7 and mth2-8I23.

Additional BACs that overlapped the interval were used to confirm certain samples. The BACs used for MYB1 transcript probing were: mth2-38K19, mth2- 49022, mth2-4L4, mth2-57D22, mth2-68B10, mth2-69F3, mth2-71F24, mth2-78L20, mth2-1G13, mth2-10P15, mth2-J16, mth2-144F11, mth2-154G18, mth2-157H22, mth2-160M23, mth2-161P23, mth2-164P8, mth2-187J11, and mth2-32AA.

4.9.3 Gel Electrophoresis

Samples were run on a 2.0% agarose minigel made with 1X Tris-Boric Acid-EDTA (TBE) buffer (89 mM Tris, 89 mM boric acid, 2 mM EDTA). Gels were then stained with ethidium bromide for visualization under UV light.

4.10 Primer Sets Utilized

Primers for PCR were generated using the FastPCR (http://www.biocenter.helsinki.fi/bi/Programs/fastpcr.htm) software program that generated primers based on the criteria of having an optimal melting temperature of 60° C and containing a purine base at both the 3' and 5' ends of the primer to help ensure specific binding. Specific melting temperature optima were altered as needed for PCR amplification from BAC templates, while RT-PCR reactions were run using a 60° C melting temperature. Additionally, the target length of the transcript was between 100-200 bp for effective RT-PCR amplification. This was later modified to be 200-300 bp to better suit both standard PCR reaction conditions (gel identification) as well as RT-PCR. Each target transcript is identified by its Affymetrix internal name and includes the length of the amplified product as well as the final melting temperature used and the melting temperature as assayed by the manufacturer, MWG (High Point, NC).

4.10.1 RT-PCR Controls for Affymetrix Sample Quality

The transcript of the constitutively expressed gene Msc27 was used as a control for the quality of the RNA product in samples prepared for Affymetrix testing (Allison et al., 1993; Csanadi et al., 1994; Dickstein et al., 2002) The gene was amplified both as a cDNA product of a reverse-transcriptase reaction as well as from genomic DNA present in the RNA sample (see Section 3.3). The product was amplified effectively at the 60° C temperature of the SYBR Green

4.10.2 Protocol.

Primers

Msc27F2 – 5’ ggaggttgagggaaagtg - T_m: 58.8° C

Msc27F2 – 5’ caccaacaaagaattgaagg - T_m: 53.2° C

Product: 311 bp

SYBR Green Protocol: 60° C – 40 seconds, 94° C - 20 seconds, 40 Cycles

RT-PCR Validation of Affymetrix Data

4.10.3 No Change Transcripts

Two transcripts were selected based on their consistent signal levels between wild type and the *nip* mutant. Both are highly expressed genes based on Affymetrix signal strength.

1) Thiolprotease

TP-F - 5'-gagcaatcgttatgcgccacg - T_m: 61.8° C

TP-R - 5'-ctgcgattgctgagaatgcccaac - T_m: 64.4° C

Amplified Length – 131 bp

SYBR Green Protocol: 60° C – 40 seconds, 94° C - 20 seconds, 40 Cycles

2) Phosphofructokinase

PFK-F - 5'-caaagtgctgttcatggagcaatg - T_m: 61.0° C

PFK-R - 5'-gtgtcttcatcggaaaccttgacc - T_m: 62.7° C

Amplified Length – 200 bp

SYBR Green Protocol: 60° C – 40 seconds, 94° C - 20 seconds, 40 Cycles

4.10.4 Upregulated Transcripts

Two transcripts were found to be highly upregulated in the *nip* samples as compared to low-levels in the wild type A17 samples.

1) hsp17.9

hsp17.9F - 5'-ccaagtgtcttcggaaccggac - T_m: 64.0° C

hsp17.9R - 5'-ggctttgaagatgtgtgcttccgg - T_m: 64.4° C

Amplified Length – 162 bp

SYBR Green Protocol: 60° C – 40 seconds, 94° C - 20 seconds, 40 Cycles

2) 17kD heat shock protein

HSP-2F - 5'-caaggcaatggctgctactcc - T_m: 61.8° C

HSP-2R - 5'-cagaaccccatcttgacaaacagc - T_m: 62.7° C

Amplified Length – 280 bp

SYBR Green Protocol: 60° C – 40 seconds, 94° C - 20 seconds, 40 Cycles

4.10.5 Downregulated Transcripts

Two transcripts were found to be highly downregulated in the *nip* samples as compared to high transcript levels in the wild type A17 samples.

1) MYB transcription factor

MYB1F – 5'-gtcaccagatacggttgctgg - T_m: 61.8° C

MYB1R - 5'-caggtaactcggccgcaatgg - T_m: 63.7° C

Amplified Length - 208 bp

SYBR Green Protocol: 60° C – 40 seconds, 94° C - 20 seconds, 40 Cycles

Also tested on BAC Templates Using BAC BLM Protocol:

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 56°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

2) Albumin 1

Alb1F - 5'-ggtggtcgaagaacacccaaac - T_m: 62.1° C

Alb1R - 5'-cagggttagagccaacctcgaag - T_m: 64.2° C

Amplified Length - 164 bp

SYBR Green Protocol: 60° C – 40 seconds, 94° C - 20 seconds, 40 Cycles

Also tested on BAC Templates Using BAC BLM Protocol:

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 57°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

4.10.6 Amplification of Transcripts from BAC Templates

The following are PCR primers used to amplify candidate PCR transcripts from BAC comprising the genetic interval surrounding *NIP*.

1) ITT

Product: 547 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 56°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

2) TrypIn

Product: 305 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 60°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

3) TC100

Product: 178 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 56°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

4) TC101

Product: 312 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 56°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

5) HSR201

Product: 201 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 56°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

6) TC102

Product: 65 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 56°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

7) Lectin Gal1

Product: 194 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 57°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

8) GAST

Product: 130 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 56°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

9) TC103

Product: 120 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 57°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

10) THIFOMT

Product: 296 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 56°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

11) MYB 2

Product: 205 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 60°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

12) DA4OACT

Product: 72 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 56°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

13) SALMSCMT

Product: 329 bp

Protocol (BAC BLM): 94°C - 5 minutes, 94°C - 30 seconds, 57°C - 30 seconds, 72°C - 30 seconds, 35 total cycles, 72°C – 5 minutes, 4°C – Hold.

REFERENCES

Affymetrix (2001) GeneChip Arrays Provide Optimal Sensitivity and Specificity for Microarray Expression Analysis. Part No. 701009 Rev 701003 https://www.affymetrix.com/support/technical/technotes/25mer_technote.pdf (22 June 2007)

Affymetrix (2004) GeneChip Expression Analysis Technical Manual. Part No. 701021 Rev. 701025 https://www.affymetrix.com/Auth/support/downloads/manuals/expression_print_manual.zip (22 June 2007)

Affymetrix (2005) Data Sheet - GeneChip Medicago Genome Array. Part No. 702030 Rev. 702031 http://www.affymetrix.com/support/technical/datasheets/medicago_datasheet.pdf (22 June 2007)

Affymetrix.com (2007) GeneChip Array Manufacturing. http://www.affymetrix.com/technology/manufacturing/index.affx (22 June 2007)

Affymetrix.com (2007) Probe Design and Selection. http://www.affymetrix.com/technology/design/index.affx (22 June 2007)

Agilent (2006) Agilent RNA 6000 Nano Kit Guide. **G2938-90034** http://cat.ucsf.edu/pdfs/Nano_RNA_Analysis.pdf (22 June 2007)

Allison LA, Kiss GB, Bauer P, Poiret M, Pierre M, Savoure A, Kondorosi A, Kondorosi E (1993) Identification of two alfalfa early nodulin genes with homology to members of the pea Enod12 gene family. Plant Molecular Biology **21:** 375-380

Boisson-Dernier A, Chabaud M, Garcia F, Becard G, Rosenberg C, Barker DG (2001) Agrobacterium rhizogenes-Transformed Roots of Medicago truncatula for the Study of Nitrogen-Fixing and Endomycorrhizal Symbiotic Associations. Molecular Plant Microbe Interactions **14:** 695-700

Boivin C, Camut S, Malpica CA, Truchet G, Rosenberg C (1990) Rhizobium meliloti Genes Encoding Catabolism of Trigonelline Are Induced under Symbiotic Conditions. Plant Cell **2:** 1157-1170

Bolstad BM (2007) RMAExpress Users Guide. http://rmaexpress.bmbolstad.com/ (22 June 2007)

Borlaug NE (2000) The Green Revolution and the Road Ahead. Norwegian Nobel Institute in Oslo **Special 30th Anniversary Nobel Lecture**

Brewin NJ (1991) Development of the Legume Root Nodule. Annual Reviews in Cell Biology **7:** 191-226

Brewin NJ (2004) Plant Cell Wall Remodelling in the Rhizobium–Legume Symbiosis. Critical Reviews in Plant Sciences **23:** 293-316

Bright L, Liang Y, Mitchell D, Harris J (2005) The LATD Gene of *Medicago truncatula* Is Required for Both Nodule and Root Development. Molecular Plant Microbe Interactions **18:** 521-532

Carpenter S, Caraco N, Correll D, Howarth R, Sharpley A, Smith V (1998) Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecological Applications **8:** 559-568

Cook DR (2001) Aeroponic Chamber for *M. truncatula*. Medicago.org http://www.medicago.org/documents/Protocols/AeroponicChamber081001.html (22 June 2007)

Cook DR, VandenBosch K, de Bruijn F, Huguet T (1997) Model Legumes Get the Nod The Plant Cell**:** 275-281

Csanadi G, Szecsi J, Kalo P, Kiss P, Endre G, Kondorosi A, Kondorosi E, Kiss GB (1994) ENOD12, an Early Nodulin Gene, Is Not Required for Nodule Formation and Efficient Nitrogen Fixation in Alfalfa. Plant Cell **6:** 201-213

Cullimore J, Dénarié J (2003) How Legumes Select Their Sweet Talking Symbionts. Science **302:** 575-578

Dickstein R, Hu X, Yang J, Ba L, Coque L, Kim D-J, Cook DR, Yeung AT (2002) Differential expression of tandemly duplicated Enod8 genes in *Medicago*. Plant Science **163:** 333-343

Dunn K, Dickstein R, Feinbaum R, Burnett BK, Peterman TK, Thoidis G, Goodman HM, Ausubel FM (1988) Developmental regulation of nodule-specific genes in alfalfa root nodules. Molecular Plant-Microbe Interactions **1:** 66-74

Eady R, Postgate J (1974) Nitrogenase. Nature **249:** 805-810

Kim J, Rees D (1992) Crystallographic structure and functional implications of the nitrogenase molybdenum-iron protein from *Azotobacter vinelandii*. Nature **360:** 553-560

Liang Y, Mitchell DM, Harris JM (2007) Abscisic acid rescues the root meristem defects of the *Medicago truncatula latd* mutant. Developmental Biology **304:** 297-307

Lullien V, Barker D, de Lajudie P, Huguet T (1987) Plant gene expression in effective and ineffective root nodules of alfalfa (*Medicago sativa*). Plant Molecular Biology **9:** 469-478

May GD, Dixon RA (2004) *Medicago truncatula*. Current Biology **14:** R180-R181

Medicago.org (2007) Sequencing Resources Project Summary Page. http://www.medicago.org/genome/about.php (22 June 2007)

Meuller O, Lightfoot S, Schroeder A (2004) RNA Integrity Number (RIN) – Standardization of RNA Quality Control. Agilent Application Note http://fgc.urmc.rochester.edu/pdf/RIN.pdf (22 June 2007)

Mitra R, Gleason C, Edwards A, Hadfield J, Downie J, Oldroyd G, Long S (2004) A Ca^{2+}/calmodulin dependent protein kinase required for symbiotic nodule development: Gene identification by transcript-based cloning. Proceedings of the National Academy of Sciences **101:** 4701-4705

Murray JD, Karas B, Sato S, Tabata S, Amyot L, Szczygloski K (2007) A Cytokinin Perception Mutant Colonized by Rhizobium in the Absence of Nodule Organogenesis. Science **315:** 101-104

Oldroyd G (2007) Nodules and Hormones. Science **315:** 52-53

Oldroyd G, Downie J (2004) Calcium, Kinases and Nodulation Signalling in Legumes. Nature Reviews Molecular Cell Biology **5:** 566-576

Oldroyd G, Long S (2003) Identification and Characterization of *Nodulation-Signaling Pathway 2*, a Gene of *Medicago truncatula* Involved in Nod Factor Signaling. Plant Physiology **131:** 1027-1032

Orme-Johnson W (1992) Nitrogenase Structure: Where to Now? Science **257:** 1639-1640

Penmetsa R, Cook D (2000) Production and Characterization of Diverse Developmental Mutants of *Medicago truncatula* Plant Physiology **123:** 1386-1397

Perret X, Staehelin C, Broughton W (2000) Molecular Basis of Symbiotic Promiscuity. Microbiology and Molecular Biology Reviews **64:** 180-201

Rabalais NN, Turner RE, Wiseman WJ (2002) Gulf of Mexico Hypoxia, a.k.a. "The Dead Zone". Annual Review of Ecology and Systematics **33:** 235-263

Schultze M, Kondorosi A (1996) The role of lipochitooligosaccharides in root nodule organogenesis and plant cell growth. Current Opinion in Genetics and Development **6:** 631-638

Schultze M, Kondorosi A (1998) REGULATION OF SYMBIOTIC ROOT NODULE DEVELOPMENT. Annual Review of Genetics **32:** 33-57

Smith B (2002) Nitrogenase Reveals Its Inner Secrets. Nature **136:** 3692-3702

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucl. Acids Res. **29:** 22-28

Vasse J, de Billy F, Camut S, Truchet G (1990) Correlation between Ultrastructural Differentiation of Bacteroids and Nitrogen Fixation in Alfalfa Nodules. Journal of Bacteriology **172:** 4295-4306

Veershlingam H, Haynes J, Penmetsa R, Cook D, Sherrier D, Dickstein R (2004) *nip*, a Symbiotic *Medicago truncatula* Mutant that Forms Noot Nodules with Aberrant Infections Threads and Plant Defense-Like Response. Plant Physiology **136:** 3692-3702

Wasson AP, Pellerone FI, Mathesius U (2006) Silencing the Flavonoid Pathway in *Medicago truncatula* Inhibits Root Nodule Formation and Prevents Auxin Transport Regulation by Rhizobia. Plant Cell **18:** 1617-1629

Young ND, Cannon SB, Sato S, Kim D, Cook DR, Town CD, Roe BA, Tabata S (2005) Sequencing the Genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiology **137:** 1174-1181

Zhang J, Subramanian S, Zhang Y, Yu O (2007) Flavone Synthases from *Medicago truncatula* Are Flavanone-2-Hydroxylases and Are Important for Nodulation. Plant Physiol. **144:** 741-751